心理学探新 2020,Vol.40,No.3,261 -268
PSYCHOLOGICAL EXPLORATION

锚测验难度水平代表性对垂直量尺化的影响研究*

叶 萌[1,2],辛 涛[2],胡卫平[1,2],孙小坚[3]
(1.陕西师范大学现代教学技术教育部重点实验室,西安 710062;2.中国基础教育质量监测协同创新中心,北京 100875;
3.西南大学数学与统计学院,重庆 400715)

摘 要:考察了锚测验难度水平对其来源测验水平的代表性对垂直量尺化的影响。采用模拟研究的方法,比较了锚测验难度等于来源测验、位于高低年级测验水平难度区间的第25百分位处及区间第50百分位处时,年级能力分布和垂直量尺特性上的参数返真结果,发现锚题难度水平高于其来源测验非但不会导致垂直量尺化结果变差,在有的情境下反而可能会提高其准确性。研究揭示人们构建垂直量尺时,可以根据内容和其他统计特征的需要对锚测验的难度水平做出适当调整。

关键词:垂直量尺化;锚测验;难度水平;锚题代表性
中图分类号:B841.2 **文献标识码**:A **文章编号**:1003 -5184(2020)03 -0261 -08

1 引言

“锚题代表性”是采用非等组锚测验(NEAT)设计进行测验链接时的一个核心问题,具有代表性的锚是否会带来更佳链接结果并非不言自明,而需实证探讨。针对统计特征代表性,目前的研究涉及等值和垂直量尺化这两个链接子领域内,锚测验的难度水平、难度范围及区分度水平。就难度水平,等值领域的研究表明锚测验具有代表性会产生更准确的等值结果(Cao,2008;Gao,Hanson,& Harris,1999,cited in Cao,2008;Marco,Petersen,& Stewart,1983,cited in Liu,Sinharay,Holland,Feigenbaum,& Curley,2011;Wei,2010)。就难度范围,等值领域的研究发现在外锚情境下,相对于难度范围代表整卷的微型锚题,用内容上对总测验具有代表性,但只包括中等难度的题目构成的锚,即midi锚做等值可能更合适,因为它得到了更高的题总相关(Sinharay & Holland,2006),等值性能和微型锚一样(Sinharay & Holland,2007),且由近期更科学的探索表明等值误差要么都小于使用微型锚时的,要么和后者非常相似(Liu,Sinharay,Holland,Feigenbaum,& Curley,2011);尽管Trierweiler,Lewis和Smith(2016)从观测锚总相关的影响因素出发进行的分析表明midi锚通常不会使锚总相关最大化,但Sinhary(2018)进一步分析了Trierweiler等(2016)考虑的多数最现实情况下的等值性能,继续为midi测验做了辩护。内锚情境下的一项IRT等值研究则表明,midi测验的等值准确性在大多数情况下和传统的微型锚测验差不多(Fitzpatrick & Skorupski,2016)。垂直量尺化领域的探索发现在外锚情境下,锚题难度范围扩大会使垂直量尺化性能和参数返真都更为准确(Chin,Kim,& Nering,2006)。就区分度水平,研究表明应用Rasch模型时区分度差异对垂直量尺化结果有影响,不同区分度水平会引入系统误差估计(Humphrey,2018)。

综上可见,测验链接领域传统的“微型版本”观念可能需要做出调整(叶萌,辛涛,2015)。就尚未涉猎的研究主题——垂直量尺化领域锚测验难度水平代表性的影响而言,如果从一个年级对中的低年级选择锚题做量尺化,当前实践是锚题的难度水平等于低年级测验水平的难度。但人们可能会质疑这样保证了锚题对低年级能力测量的准确性,却可能牺牲了高年级的准确性(叶萌,辛涛,2015)。另一方面,锚题构建实践很可能会出现锚测验其他特征都符合要求,但难度水平和来源测验不同的情况,那么这样的锚测验能否用来构建垂直量尺?研究聚焦于探讨垂直量尺化领域中,锚测验的难度水平对其来源测验的代表性不同,是否会导致不同的垂直量尺化结果。由于既有研究普遍显示测验长度可能对链接情境较敏感,且量尺标定方法的选择对量尺化结果有影响,因此研究也将测验长度和参数标定方法作为探索因素。

2 研究方法

研究采用模拟研究的方法范式。

2.1 数据来源和模拟数据

假设有5个需要进行垂直量尺化的测验水平,分别覆盖了3至7年级的教学内容,中间年级5年级是量尺化基准水平。假设各年级均有2000考生,能力分布均为正态分布,均值由低到高分别为 -1、

* **基金项目**:教育部人文社会科学研究西部和边疆地区青年基金项目(15XJC190003),国家社会科学基金青年项目(17CTJ014)。
通讯作者:辛涛,E -mail:xintao@bnu.edu.cn。

-0.5、0、0.5 和 1，标准差为 1。

采用内锚 NEAT 设计进行量尺链接。设测验水平的题目数有两个水平：50 和 30，分别代表长测验和较短测验。锚测验的构建方式为从相邻测验水平对中的较低水平里，选择也适合较高年级学生作答的题目。锚题长度为测验水平的 20%（针对 50 题情境，有 10 个锚题）或 25%（针对 30 题情境，有 8 个锚题）。根据这些设定对 O'Neil(2010) 的研究中时间 1 上的 3 参数 logistic 模型（3PLM）题目参数（每年级 70 题，且参数未经量尺化）进行筛选和微调，分别得到量尺化的各测验水平题目参数集和锚题参数集。研究基线条件为锚测验难度等于其来源测验水平难度。基线条件下各测验水平和锚测验的题目难度参数分布特征分别如表 1 和表 2 所示。

表 1 基线条件下各测验水平的题目难度参数分布特征

	50 题测验		30 题测验	
	M	*SD*	*M*	*SD*
G3	-0.953	0.662	-0.973	0.730
G4	-0.514	0.896	-0.511	1.024
G5	-0.022	0.994	-0.041	1.164
G6	0.611	1.265	0.497	1.333
G7	0.949	1.033	1.011	1.186

注：*M* = 均值，*SD* = 标准差，下同。

采取锚题集整体上移的方式，将锚测验的难度水平上移两次，令其分别位于高低年级难度区间的第 25 百分位处和第 50 百分位处，得到一个实验刺激：锚测验难度变化，并（出于简便起见，且考虑到锚测验是从低年级抽取的）将其三个水平命名为："+0%"（无变化，即基线条件）、"+25%"（上移 25 百分位，位于高低年级难度区间的第 25 百分位处），及"+50%"（上移 50 百分位，位于高低年级难度区间的第 50 百分位处）。这样，比如，对于 50 题测验，在 +25% 水平上，3、4 年级的锚测验中所有锚题的难度都将上移 0.11*，锚测验均值也相应地（由基线的 -0.935）变为 -0.823。

针对以上设置生成模拟数据，各种研究条件下都执行 50 个复本。

表 2 基线条件下各测验水平间锚测验的题目难度参数分布特征

	50 题测验		30 题测验	
	M	*SD*	*M*	*SD*
G34	-0.935	0.864	-0.914	0.949
G45	-0.487	0.932	-0.509	1.051
G56	-0.016	1.142	-0.004	1.199
G67	0.609	1.429	0.576	1.441

2.2 量尺化方法和程序

使用 BILOG - MG 程序进行参数估计，其中能力参数估计方法为 EAP。使用的量尺转换方法包括同时标定和 4 种分别标定——用 Stocking & Lord 转换的分别标定（以下简称"SL 转换"）、用 Habera 转换的分别标定（以下简称"Habera 转换"）、用均值 - 均值转换的分别标定（以下简称"MM 转换"），及用均值 - 标准差转换的分别标定（以下简称"MS 转换"）。量尺转换的计算机程序为 ST 2.0。

综上，研究是一个 2（测验长度：50 题和 30 题）× 3（锚测验难度变化：+0%、+25%，及 +50%）×5（量尺转换方法：同时标定、SL 转换、Habera 转换、MM 转换，及 MS 转换）的设计，共含 30 种研究条件。

2.3 评价标准

考虑到实践者不仅关注垂直量尺特性所表征的学业发展，也会关注各个年级的能力分布，研究同时根据这两方面参数的返真度来评价垂直量尺化结果的准确性。年级能力分布表征为特定年级能力分布的均值和标准差（SD）。垂直量尺的特性主要包括三个指标：跨年级增长（相邻年级量尺分数均值之差），跨年级变异（相邻年级量尺分数分布的标准差之差），以及年级分布的分隔（separation）。年级分布的分隔用效应值来表示，数学表达式为：

$$effectsize = \frac{\bar{X}_{upper} - \bar{X}_{lower}}{\sqrt{\frac{S^2 + S^2}{2}}}。\quad (1)$$

其中 $\bar{X}_{upper}$ 和 $\bar{X}_{lower}$ 分别表示高年级和低年级的量尺分数均值，S^2 和 S^2 分别表示高年级和低年级量尺分数分布的方差。针对这 5 个参数，计算偏差和误差的均方根（RMSE）来表明其返真度，二者的临界值分别设为 0.1 和 0.2。

3 研究结果

3.1 能力均值

表 3 呈现的是能力均值估计的返真度。

在 50 题测验的情境下，除 +50% 条件下采用 MS 转换得到的 7 年级能力均值估计外，不论锚题难度水平和量尺化方法如何组合，各年级的能力均值估计都较为准确，且越靠近基准年级准确度越高。另外，同样是远离基准年级，3 年级的估计比 7 年级更为准确。从估计偏差来看，同时标定低估了各年级能力均值；分别标定则倾向于低估低年级的能力均值，高估高年级的能力均值。在误差的量上，MM 转换和 MS 转换带来的估计误差更大，其他三种标

* 该研究整体数据均保留三位小数点。但 O'Neil(2010) 的源数据只保留两位小数点，因此在该研究的数据生成阶段，为了使锚测验格式和其他题目保持一致，仍然遵循 O'Neil 的惯例，锚题难度上移量保留的是两位小数点。

定方法性能接近。

表3　不同量尺化条件下能力均值估计的返真度

测验长度	标定方法	锚测验难度变化	3年级		4年级		5年级		6年级		7年级	
			偏差	RMSE	偏差	RMSE	偏差	RMSE	偏差	RMSE	偏差	RMSE
50	conc	+0%	−0.051	0.071	−0.012	0.036	−0.004	0.021	−0.023	0.035	−0.062	0.073
		+25%	−0.049	0.071	−0.010	0.037	−0.003	0.021	−0.020	0.033	−0.061	0.072
		+50%	−0.047	0.070	−0.009	0.037	−0.004	0.021	−0.019	0.033	−0.056	0.070
	H	+0%	−0.048	0.067	−0.031	0.047	−0.006	0.021	0.017	0.033	0.049	0.070
		+25%	−0.041	0.064	−0.030	0.046	−0.006	0.022	0.019	0.036	0.051	0.071
		+50%	−0.037	0.060	−0.029	0.042	−0.007	0.021	0.020	0.036	0.053	0.074
	SL	+0%	−0.044	0.065	−0.033	0.049	−0.006	0.021	0.014	0.032	0.048	0.069
		+25%	−0.040	0.066	−0.032	0.049	−0.006	0.022	0.015	0.034	0.050	0.069
		+50%	−0.038	0.062	−0.031	0.046	−0.007	0.021	0.015	0.035	0.051	0.070
	MM	+0%	−0.054	0.084	−0.025	0.051	−0.006	0.021	0.009	0.040	0.073	0.108
		+25%	−0.051	0.085	−0.023	0.051	−0.006	0.022	0.012	0.043	0.079	0.112
		+50%	−0.050	0.088	−0.025	0.051	−0.007	0.021	0.015	0.048	0.087	0.118
	MS	+0%	−0.044	0.071	−0.024	0.051	−0.006	0.021	0.011	0.041	0.076	0.118
		+25%	−0.035	0.069	−0.020	0.049	−0.006	0.022	0.016	0.045	0.089	0.129
		+50%	−0.036	0.071	−0.022	0.046	−0.007	0.021	0.018	0.050	0.099	0.138
30	conc	+0%	0.090	0.097	0.059	0.065	0.004	0.021	−0.079	0.084	−0.162	0.167
		+25%	0.093	0.104	0.059	0.066	0.008	0.022	−0.071	0.076	−0.163	0.165
		+50%	0.100	0.110	0.064	0.070	0.008	0.022	−0.073	0.077	−0.162	0.165
	H	+0%	−0.025	0.052	−0.014	0.038	0.002	0.020	0.031	0.048	0.074	0.102
		+25%	−0.024	0.054	−0.016	0.036	0.006	0.020	0.034	0.047	0.066	0.086
		+50%	−0.012	0.048	−0.009	0.032	0.006	0.020	0.034	0.045	0.071	0.087
	SL	+0%	−0.019	0.053	−0.010	0.037	0.002	0.020	0.028	0.045	0.072	0.098
		+25%	−0.013	0.062	−0.013	0.036	0.006	0.020	0.032	0.047	0.056	0.079
		+50%	−0.004	0.058	−0.008	0.032	0.006	0.020	0.031	0.045	0.058	0.078
	MM	+0%	−0.027	0.069	−0.014	0.044	0.002	0.020	0.036	0.063	0.099	0.141
		+25%	−0.027	0.077	−0.016	0.047	0.006	0.020	0.042	0.064	0.093	0.123
		+50%	−0.016	0.078	−0.006	0.046	0.006	0.020	0.046	0.069	0.104	0.132
	MS	+0%	−0.026	0.064	−0.014	0.044	0.002	0.020	0.036	0.067	0.100	0.143
		+25%	−0.026	0.069	−0.015	0.044	0.006	0.020	0.045	0.069	0.100	0.134
		+50%	−0.016	0.064	−0.008	0.040	0.006	0.020	0.049	0.074	0.115	0.148

注：conc = 同时标定；H = 使用 Haebara 转换的分别标定；SL = 使用 Stocking & Lord 转换的分别标定；MM = 使用均值 − 均值转换的分别标定；MS = 使用均值 − 标准差转换的分别标定；下同。

对于长测验，锚测验难度变化对能力均值估计的影响整体并未达到使返真度显著扭曲的程度。不过具体影响模式呈现出了一些规律性。首先，偏差和 RMSE 随着锚题难度上移而产生的变化的方向整体相同，不过 RMSE 变化的程度小于偏差。第二，综合偏差和 RMSE，使用同时标定时，锚题难度上移使各个年级能力均值的估计都倾向于更为准确。使用分别标定时，锚题难度上移倾向于使低年级的能力均值估计准确性提高，而使高年级的估计准确性有所降低，对基准年级则没什么影响。不过，使用 MS 转换时，低年级的估计偏差变化不稳定，上移 25% 和上移 50% 这两种情境似乎没什么区别。第三，越靠近基准年级的年级，能力估计随锚题难度上移的变化幅度越小；反之，越接近量尺两端的年级，能力估计受锚题难度变化的影响越大，尤其是 7 年级。第四，MM 和 MS 方法得到的量尺两端年级（尤其是高年级）的能力估计更容易受到锚题难度变化的影响，尤其是 MS 方法，其他三种方法没有明显区别。

30 题测验情境下的量尺化结果和长测验有所不同。标定方法自身的量尺化性能显示，对于非基准年级，同时标定系统高估了低年级能力均值而低估了高年级能力均值（这不同于长测验情境），几种分别标定则系统低估低年级能力而高估高年级能力（和长测验情境相同）；对于基准年级，所有标定方法都高估其平均能力（和长测验相反）。在估计偏移程度上，MM 和 MS 转换使 7 年级能力均值估计显著系统有偏；同时标定同时使 3 年级和 7 年级的估计出现显著偏差，且偏离程度较大；SL 转换和 Hae-

bara 转换则几乎在所有条件下都能很准确地估计能力均值。

较短测验情境下锚题难度变化对能力均值估计的影响同样不大。就影响的规律性，长测验情境下的后两个规律在较短测验情境下得到了复制，前两个则并未在此发现。具体说来，偏差和 RMSE 随着锚题难度上移而产生的变化在程度上似乎没有明确的大小之分，在方向上也没有固定的一致性模式（除使用同时标定时变化的方向一致外）。与之相关，表中结果也未能表明均值估计受到了锚题难度变化的稳定影响，看起来它在不同锚题难度水平下的变化更像是随机变异。

表 4 不同量尺化条件下能力分布 SD 估计的返真度

测验长度	标定方法	锚测验难度变化	3 年级		4 年级		5 年级		6 年级		7 年级	
			偏差	RMSE	偏差	RMSE	偏差	RMSE	偏差	RMSE	偏差	RMSE
50	conc	+0%	0.132	0.138	0.070	0.079	0.001	0.019	−0.057	0.064	−0.064	0.073
		+25%	0.121	0.126	0.064	0.074	0.001	0.019	−0.057	0.064	−0.060	0.071
		+50%	0.115	0.121	0.059	0.069	0.001	0.019	−0.053	0.061	−0.055	0.067
	H	+0%	0.023	0.048	0.027	0.041	0.021	0.027	0.015	0.035	0.033	0.060
		+25%	0.022	0.048	0.019	0.037	0.020	0.026	0.014	0.032	0.036	0.062
		+50%	0.024	0.048	0.013	0.034	0.020	0.026	0.016	0.035	0.041	0.069
	SL	+0%	−0.011	0.042	0.016	0.039	0.021	0.027	0.014	0.038	0.018	0.056
		+25%	−0.007	0.043	0.014	0.037	0.020	0.026	0.014	0.035	0.021	0.055
		+50%	0.002	0.039	0.011	0.034	0.020	0.026	0.016	0.035	0.025	0.060
	MM	+0%	0.060	0.098	0.033	0.058	0.021	0.027	0.010	0.052	0.052	0.097
		+25%	0.057	0.098	0.026	0.058	0.020	0.026	0.007	0.052	0.058	0.103
		+50%	0.045	0.096	0.021	0.054	0.020	0.026	0.013	0.054	0.067	0.104
	MS	+0%	0.010	0.055	0.016	0.041	0.021	0.027	0.014	0.047	0.049	0.086
		+25%	0.010	0.054	0.008	0.040	0.020	0.026	0.019	0.045	0.065	0.098
		+50%	0.017	0.052	0.007	0.038	0.020	0.026	0.028	0.052	0.079	0.115
30	conc	+0%	0.005	0.021	−0.025	0.031	−0.058	0.060	−0.072	0.075	−0.067	0.072
		+25%	−0.008	0.032	−0.035	0.042	−0.064	0.067	−0.077	0.081	−0.072	0.078
		+50%	−0.011	0.031	−0.039	0.045	−0.066	0.069	−0.074	0.078	−0.068	0.074
	H	+0%	−0.025	0.041	−0.026	0.037	−0.027	0.030	−0.020	0.036	−0.004	0.053
		+25%	−0.034	0.058	−0.032	0.047	−0.030	0.033	−0.024	0.048	−0.006	0.060
		+50%	−0.033	0.057	−0.037	0.051	−0.030	0.033	−0.018	0.045	0.002	0.060
	SL	+0%	−0.034	0.050	−0.024	0.037	−0.027	0.030	−0.026	0.045	−0.010	0.053
		+25%	−0.048	0.072	−0.031	0.053	−0.030	0.033	−0.034	0.052	−0.019	0.061
		+50%	−0.045	0.068	−0.035	0.053	−0.030	0.033	−0.030	0.050	−0.014	0.058
	MM	+0%	−0.015	0.065	−0.022	0.061	−0.027	0.030	−0.018	0.051	−0.004	0.082
		+25%	−0.022	0.075	−0.032	0.062	−0.030	0.033	−0.018	0.069	0.010	0.087
		+50%	−0.020	0.085	−0.045	0.070	−0.030	0.033	−0.010	0.065	0.022	0.086
	MS	+0%	−0.025	0.053	−0.026	0.040	−0.027	0.030	−0.019	0.043	0.000	0.065
		+25%	−0.034	0.062	−0.030	0.049	−0.030	0.033	−0.011	0.051	0.014	0.075
		+50%	−0.036	0.060	−0.037	0.053	−0.030	0.033	0.003	0.056	0.034	0.089

3.2 能力分布 SD

表 4 呈现了能力分布 SD 估计的返真度。

对于 50 题测验，除同时标定得到的 3 年级能力均值估计外，在各种研究条件下，各年级 SD 估计都较为准确。估计的 RMSE 越靠近基准年级越低；估计偏差上只有同时标定显示出了这样的趋势，对于分别标定，最小估计偏移出现在哪个年级较为随机。比较标定方法来看，同时标定倾向于高估低年级的 SD 而低估高年级的 SD；分别标定则高估了所有年级的 SD。RMSE 显示 SL 转换的估计误差最小，其次是 Haebara 转换；同时标定的性能最差。最后，锚题难度变化对 SD 估计也没有实质性影响。在影响模式的规律性上，使用同时标定时，锚题难度上移使各年级 SD 的估计都倾向于更为准确。

30 题测验情境下的结果显示，在大多数条件下，垂直量尺化过程都会系统低估 SD 真值，但没有显著的估计误差。不同标定方法得到的结果没有明显的优劣之分。锚题难度变化对 SD 估计的影响也不大，且没有什么规律性影响模式。

表5　不同量尺化条件下跨年级增长估计的返真度

测验长度	标定方法	锚测验难度变化	3至4年级		4至5年级		5至6年级		6至7年级	
			偏差	RMSE	偏差	RMSE	偏差	RMSE	偏差	RMSE
50	conc	+0%	0.040	0.051	0.007	0.029	−0.018	0.026	−0.040	0.046
		+25%	0.039	0.051	0.007	0.031	−0.017	0.025	−0.041	0.047
		+50%	0.038	0.051	0.005	0.033	−0.014	0.023	−0.038	0.046
	H	+0%	0.017	0.035	0.025	0.036	0.023	0.030	0.032	0.050
		+25%	0.011	0.033	0.024	0.035	0.025	0.034	0.032	0.048
		+50%	0.009	0.031	0.021	0.032	0.027	0.036	0.034	0.049
	SL	+0%	0.011	0.032	0.027	0.039	0.020	0.029	0.033	0.054
		+25%	0.008	0.031	0.027	0.039	0.021	0.031	0.034	0.051
		+50%	0.006	0.029	0.024	0.038	0.023	0.033	0.035	0.051
	MM	+0%	0.029	0.051	0.019	0.040	0.015	0.039	0.064	0.092
		+25%	0.029	0.053	0.017	0.040	0.017	0.043	0.067	0.094
		+50%	0.025	0.052	0.018	0.040	0.022	0.047	0.072	0.097
	MS	+0%	0.019	0.044	0.018	0.040	0.017	0.038	0.065	0.098
		+25%	0.015	0.044	0.015	0.038	0.022	0.045	0.073	0.106
		+50%	0.014	0.044	0.014	0.037	0.026	0.050	0.081	0.113
30	conc	+0%	−0.031	0.039	−0.055	0.058	−0.083	0.086	−0.083	0.087
		+25%	−0.034	0.043	−0.051	0.056	−0.080	0.082	−0.091	0.094
		+50%	−0.036	0.044	−0.056	0.060	−0.081	0.083	−0.089	0.092
	H	+0%	0.011	0.026	0.016	0.031	0.029	0.041	0.043	0.066
		+25%	0.008	0.035	0.022	0.036	0.028	0.041	0.032	0.051
		+50%	0.003	0.032	0.015	0.032	0.028	0.038	0.038	0.053
	SL	+0%	0.009	0.028	0.011	0.031	0.026	0.038	0.044	0.069
		+25%	0.000	0.038	0.019	0.037	0.025	0.039	0.024	0.049
		+50%	−0.004	0.037	0.014	0.034	0.025	0.038	0.027	0.049
	MM	+0%	0.013	0.040	0.016	0.037	0.035	0.056	0.062	0.103
		+25%	0.011	0.047	0.022	0.047	0.036	0.059	0.051	0.076
		+50%	0.009	0.051	0.012	0.044	0.040	0.065	0.058	0.081
	MS	+0%	0.011	0.039	0.016	0.037	0.034	0.060	0.065	0.103
		+25%	0.011	0.047	0.021	0.045	0.039	0.065	0.055	0.082
		+50%	0.008	0.045	0.014	0.041	0.043	0.071	0.066	0.091

3.3　跨年级增长

表5显示，不论测验长短，不论锚题难度水平如何变化，在各种量尺化方法下，跨年级增长的估计都较准确，且低年级增长的估计准确性高于高年级。

在长测验情境下，同时标定高估了量尺低端的增长而低估高端的增长，分别标定在整个量尺上都高估了跨年级增长。综合偏差和RMSE来看，SL和Haebara转换都能很好地估计整个量尺上的跨年级增长，MM和MS转换对6、7年级的增长估计相对较差。使用同时标定时，锚题难度上移会使跨年级增长估计更准确；使用分别标定时，锚题难度上移会使量尺低年级端的跨年级增长估计更准确，而使高年级端的增长估计准确性降低。

30题测验情境下，同时标定系统低估了跨年级增长，分别标定则系统高估了该参数。从量尺化性能上看，同时标定最差，SL转换最佳。锚题难度变化对跨年级增长估计似乎没什么影响，不同难度水平下的跨年级增长量呈随机性。

表6　不同量尺化条件下跨年级变异估计的返真度

测验长度	标定方法	锚测验难度变化	3至4年级		4至5年级		5至6年级		6至7年级	
			偏差	RMSE	偏差	RMSE	偏差	RMSE	偏差	RMSE
50	conc	+0%	−0.062	0.071	−0.069	0.076	−0.058	0.062	−0.007	0.033
		+25%	−0.056	0.065	−0.063	0.070	−0.058	0.062	−0.004	0.032
		+50%	−0.057	0.066	−0.057	0.065	−0.054	0.059	−0.002	0.030
	H	+0%	0.004	0.027	−0.006	0.028	−0.006	0.025	0.018	0.044
		+25%	−0.003	0.028	0.001	0.027	−0.006	0.025	0.022	0.045
		+50%	−0.011	0.030	0.007	0.028	−0.004	0.027	0.025	0.049
	SL	+0%	0.027	0.037	0.005	0.033	−0.007	0.031	0.005	0.042
		+25%	0.020	0.036	0.007	0.031	−0.006	0.029	0.008	0.041
		+50%	0.009	0.028	0.009	0.030	−0.004	0.027	0.008	0.044

续表 6

测验长度	标定方法	锚测验难度变化	3 至 4 年级		4 至 5 年级		5 至 6 年级		6 至 7 年级	
			偏差	RMSE	偏差	RMSE	偏差	RMSE	偏差	RMSE
	MM	+0%	-0.027	0.057	-0.012	0.049	-0.011	0.048	0.042	0.077
		+25%	-0.030	0.062	-0.006	0.048	-0.013	0.050	0.050	0.082
		+50%	-0.023	0.064	-0.001	0.048	-0.007	0.050	0.054	0.081
	MS	+0%	0.006	0.036	0.005	0.035	-0.007	0.040	0.035	0.068
		+25%	-0.002	0.039	0.012	0.035	-0.001	0.036	0.046	0.076
		+50%	-0.010	0.034	0.013	0.037	0.008	0.041	0.051	0.085
30	conc	+0%	-0.030	0.035	-0.034	0.038	-0.014	0.021	0.005	0.020
		+25%	-0.026	0.035	-0.030	0.035	-0.012	0.022	0.005	0.021
		+50%	-0.028	0.036	-0.027	0.032	-0.008	0.020	0.005	0.022
	H	+0%	-0.001	0.025	-0.002	0.028	0.008	0.029	0.016	0.041
		+25%	0.002	0.035	0.002	0.030	0.006	0.039	0.018	0.050
		+50%	-0.004	0.034	0.007	0.031	0.012	0.040	0.021	0.053
	SL	+0%	0.010	0.026	-0.004	0.030	0.001	0.034	0.016	0.048
		+25%	0.017	0.036	0.001	0.038	-0.004	0.033	0.015	0.047
		+50%	0.010	0.034	0.005	0.035	0.000	0.036	0.016	0.047
	MM	+0%	-0.007	0.055	-0.005	0.057	0.010	0.049	0.013	0.065
		+25%	-0.010	0.058	0.002	0.050	0.012	0.067	0.029	0.077
		+50%	-0.025	0.066	0.015	0.051	0.021	0.067	0.032	0.075
	MS	+0%	-0.001	0.032	-0.001	0.032	0.009	0.040	0.019	0.056
		+25%	0.004	0.038	0.000	0.036	0.020	0.050	0.024	0.057
		+50%	-0.002	0.037	0.007	0.035	0.033	0.062	0.031	0.067

3.4 跨年级变异

跨年级变异估计的返真度如表 6 所示，其在各研究条件下都没有显著扭曲。

在长测验情境下，同时标定会低估跨年级变异，其他标定方法的得到的结果则没有稳定一致的偏移方向。从误差的量上看，SL 转换对跨年级变异的估计性能最佳，其次是 Haebara 转换，同时标定性能最差。使用同时标定时，随着锚题难度上移，跨年级变异的估计趋向于更为准确，使用其他参数标定方法时则没有出现这种效应。

使用较短测验时，在基线条件下，各标定方法得到的各个跨年级变异估计没有一致的偏移方向，整体来看出现低估的情况比较多；各方法得到的误差的量也没有多少差异。而且，锚题难度变化对跨年级变异估计似乎也没什么影响。

表 7 不同量尺化条件下年级间效应值估计的返真度

测验长度	标定方法	锚测验难度变化	3 至 4 年级		4 至 5 年级		5 至 6 年级		6 至 7 年级	
			偏差	RMSE	偏差	RMSE	偏差	RMSE	偏差	RMSE
50	conc	+0%	-0.011	0.027	-0.011	0.026	-0.005	0.019	-0.010	0.027
		+25%	-0.007	0.028	-0.009	0.027	-0.004	0.019	-0.012	0.027
		+50%	-0.006	0.028	-0.010	0.030	-0.002	0.018	-0.011	0.028
	H	+0%	0.004	0.023	0.012	0.027	0.014	0.022	0.019	0.036
		+25%	0.001	0.022	0.014	0.026	0.016	0.026	0.019	0.035
		+50%	-0.001	0.022	0.013	0.026	0.018	0.028	0.019	0.035
	SL	+0%	0.009	0.023	0.017	0.032	0.012	0.023	0.024	0.039
		+25%	0.006	0.022	0.018	0.031	0.012	0.026	0.024	0.037
		+50%	0.003	0.021	0.016	0.031	0.014	0.026	0.024	0.037
	MM	+0%	0.006	0.029	0.005	0.031	0.007	0.031	0.047	0.071
		+25%	0.007	0.030	0.005	0.031	0.010	0.037	0.049	0.075
		+50%	0.008	0.030	0.007	0.032	0.014	0.041	0.050	0.077
	MS	+0%	0.012	0.032	0.009	0.035	0.008	0.031	0.046	0.071
		+25%	0.010	0.032	0.008	0.032	0.011	0.037	0.048	0.074

续表 7

测验长度	标定方法	锚测验难度变化	3至4年级		4至5年级		5至6年级		6至7年级	
			偏差	RMSE	偏差	RMSE	偏差	RMSE	偏差	RMSE
30	conc	+50%	0.008	0.030	0.008	0.033	0.013	0.040	0.049	0.074
		+0%	−0.027	0.034	−0.035	0.040	−0.054	0.058	−0.052	0.059
		+25%	−0.024	0.032	−0.028	0.036	−0.048	0.051	−0.058	0.061
		+50%	−0.024	0.032	−0.032	0.039	−0.050	0.053	−0.058	0.060
	H	+0%	0.025	0.033	0.030	0.038	0.042	0.048	0.049	0.065
		+25%	0.026	0.038	0.038	0.048	0.043	0.050	0.040	0.050
		+50%	0.022	0.033	0.032	0.042	0.041	0.047	0.042	0.052
	SL	+0%	0.024	0.034	0.024	0.036	0.041	0.048	0.053	0.072
		+25%	0.021	0.037	0.035	0.045	0.043	0.051	0.039	0.052
		+50%	0.017	0.032	0.031	0.041	0.042	0.049	0.039	0.053
	MM	+0%	0.024	0.042	0.029	0.042	0.047	0.063	0.067	0.098
		+25%	0.026	0.046	0.037	0.053	0.049	0.065	0.054	0.076
		+50%	0.027	0.046	0.031	0.050	0.051	0.071	0.056	0.079
	MS	+0%	0.025	0.041	0.030	0.044	0.046	0.064	0.068	0.098
		+25%	0.027	0.047	0.037	0.055	0.049	0.067	0.054	0.076
		+50%	0.027	0.046	0.031	0.050	0.050	0.069	0.056	0.079

3.5　效应值

表7显示，在各种研究条件下，年级间效应值都能得到准确估计。

在长测验情境下，年级间效应值被同时标定低估，而被分别标定高估。就估计误差的大小而言，同时标定性能最佳，其次是Haebara转换，MS和MM转换则表现较差。锚测验难度变化似乎对MM转换产生了影响，随着难度上移，MM转换得到的效应值估计倾向于更偏离其真值。较短测验情境下各标定方法的估计偏移方向和长测验相同。同时标定和MS转换的返真度相对较差，SL转换和Haebara转换则得到了很好的返真。锚题难度变化对年级间效应值估计似乎没什么影响。

4　讨论

研究发现锚题测验难度水平高于其来源测验非但未必会导致量尺化结果变差，而且有时可能会提高其准确性。具体来说，在长测验情境下，锚测验难度上移会使同时标定得到的量尺化结果（效应值除外）更为准确，使分别标定在估计能力均值和跨年级增长这两个参数时，低年级的准确性提高而高年级的准确性降低；在较短测验情境下不同锚题难度之间的量尺化结果则呈现出随机性。研究揭示内锚情境下锚测验难度水平对量尺化没什么影响，进行垂直量尺化时完全可根据内容和其他统计特征的需要对锚测验的难度水平做出适当调整。

该研究的结果和等值领域的锚测验难度水平研究结果有所不同。这种差异性可能缘于两个领域中锚测验在来源和施测上的不同。等值中的锚题来源于并施测给平行试卷，其难度水平偏离试卷难度意味着锚测验平均难度和考生平均能力不匹配，这样基于锚测验的考生能力估计误差增大。而在该研究的垂直量尺化语境下，锚题来源于一对测验水平中的低年级水平，并同时施测给高年级水平，这意味着锚测验完全代表低年级水平，完全不代表高年级水平，而锚题难度上移令锚测验对两个年级都部分代表。这样，从误差累积的角度来讲，代表和不代表方案的考生能力估计误差，及进一步的能力差异估计误差，孰大孰小似乎难以定论。另一方面，研究采用内锚设计意味着，锚测验难度变化也使相应测验水平的难度发生变化，而由于众多年级特异性题目的稀释作用，这种变化的程度相较于锚测验的要小得多，因此锚测验变化即便有影响也不大。在外锚语境下进行探索有助于更进一步阐释锚测验难度水平代表性的效应。另外，鉴于本研究的结论与Fitzpatrick和Skorupski（2016）在内锚情境下实施的等值锚测验难度范围研究结论一致，下一步值得思考锚测验施测方式是否对测验链接性能有影响。

参数标定方法的相对性能似乎不受锚测验难度水平的影响。研究表明在有5个待量尺化年级时，同时标定的性能不够稳定，不论测验长短都是如此。这支持了Chin等（2006）在同一链接子领域的研究发现。特征曲线估计的性能非常稳定，不论测验长短，它们对于各个参数都能做到很好地估计；积距转换仍然性能不佳。这些结果和测验链接领域的普遍发现一致。研究还发现锚题难度上移在长测验情境下导致了部分规律性的结果，而较短测验情境下的结果则基本呈随机性。两种长度下各测验水平及锚测验的偏度分析表明各题目集均服从正态分布，表1和表2也显示两种长度下锚测验和其来源测验水平的统计特征及二者的对比具有一致性，原始题目的考察也排除了极端题影响。关于该结果的解释仍需进一步研究探索。

研究模拟的是由二级计分的单选题构成的单维测验，实践中可能会应用更复杂的测验形式，因此该研究只是垂直量尺化中的锚测验难度水平代表性这个问题的研究起点。另外，考虑到实践价值，该研究

没有考察特别短的测验，比如小于 20 题的。但有理由推断在这种测验长度下得到的锚题代表性效应可能会有所不同，因此如果实践使用很短的测验，建议量尺化实施者另行研究该问题。

参考文献

叶萌，辛涛.（2015）. 测验链接中的锚题代表性研究. *心理科学*，*38*（1），209 - 215.

Cao，Y.（2008）. *Mixed - format test equating effects of test dimensionality and common - item sets*（Unpublished doctorial dissertation）. University of Maryland，Maryland，US.

Gao，X - H.，Hanson，B. A.，& Harris，D. J.（1999）. *Effect of using different common item sets under the common item non - equivalent groups design*. Paper presented at the Annual Meeting of the American Educational Research Association，Montreal，Canada. In Y. Cao（Ed.），*Mixed - format test equating effects of test dimensionality and common - item sets*（Unpublished doctoral dissertation）. University of Maryland，Maryland，US.

Chin，T. Y.，Kim，W.，& Nering，M. L.（2006）. *Five statistical factors that influence IRT vertical scaling*. Paper presented at the Annual Meeting of the National Council on Measurement in Education，San Francisco，CA.

Fitzpatrick，J.，& Skorupski，W. P.（2016）. Equating with miditests using IRT. *Journal of Educational Measurement*，*53*（2），172 - 189.

Humphrey，S. N.（2018）. The impact of levels of discrimination on vertical equating in the Rasch model. *Journal of Applied Measurement*，*19*（3），216 - 228.

Liu，J. S.，Sinharay，S.，Holland，P.，Feigenbaum，M.，& Curley，E.（2011）. Observed score equating using a mini - version anchor and an anchor with less spread of difficulty：A comparison study. *Educational and Psychological Measurement*，*71*（2），346 - 361.

Marco，G. L.，Petersen，N. S.，& Stewart，E. E.（1983）. *A large - scale evaluation of linear and curvilinear score equating models*（ETS Research Memorandum No. RM - 83 - 2）. Princeton，NJ：Educational Testing Service. Cited in J. S. Liu，S. Sinharay，P. Holland，M. Feigenbaum，& E. Curley（Eds.），Observed score equating using a mini - version anchor and an anchor with less spread of difficulty：A comparison study. *Educational and Psychological Measurement*，*71*（2），346 - 361.

O'Neil，T. P.（2010）. *Maintenance of vertical scales under conditions of item parameter drift and Rasch model - data misfit*（Unpublished doctorial dissertation）. University of Massachusetts - Amherst，Massachusetts，US.

Sinharay，S.，& Holland，P. W.（2006）. *The correlation between the scores of a test and an anchor test*（ETS RR - 06 - 04）. Princeton，NJ：Educational Testing Service.

Sinharay，S.，& Holland，P. W.（2007）. Is it necessary to make anchor tests mini - versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement*，*44*（3），249 - 275.

Sinharay，S.（2018）. On the choice of anchor tests in equating. *Educational Measurement - Issues and Practice*，*37*（2），64 - 69.

Trierweiler，T. J.，Lewis，C.，& Smith，R. L.（2016）. Further study of the choice of anchor tests in equating. *Journal of Educational Measurement*，*53*（4），498 - 518.

Wei，H.（2010）. *Impact of non - representative anchor items on scale stability*. Paper presented at the Annual Meeting of the National Council on Measurement in Education，Colorado，US.

The Impact of the Representativeness of Difficulty Level of Anchor Test on Vertical Scaling

Ye Meng[1,2]，Xin Tao[2]，Hu Weiping[1,2]，Sun Xiaojian[3]

（1. Key Laboratory of Modern Teaching Technology，Ministry of Education，Shaanxi Normal University，Xi'an 710062；
2. China Basic Education Quality Monitoring and Cooperative Innovation Center，Beijing 100875；
3. School of Mathematics and Statistics，Southwest University，Chongqing 400715）

Abstract：The influence of the representativeness of difficulty level of anchor test to its source test level on vertical scaling was investigated. Using the method of simulation，it compared recovery of the parameters in grade ability distribution and features of the vertical scale among conditions of difficulty level of anchor test equal to the source test，at the 25th percentile of the difficulty interval composed of the upper and lower test levels and at the 50th percentile of the interval. The results indicated that rather than deteriorating the results of vertical scaling，anchor test with difficulty level higher than that of its source test may improve its accuracy in some situations. The research revealed that the difficulty level of anchor test can be adjusted appropriately according to the needs of content and other statistical characteristics when vertical scale is constructed.

Key words：vertical scale；anchor test；difficulty level；representativeness of anchor test