心理与行为研究 2026，24(2)：151~160
Studies of Psychology and Behavior
DOI：10.12139/j.1672-0628.2026.02.002

两两比较在汉语文本难度评估中的应用 *

杨官豆豆[1] 谭静文[1,2] 刘苗苗[1,3] 李 虹[1]

(1 北京师范大学心理学部，应用实验心理北京市重点实验室，心理学国家级实验教学示范中心(北京师范大学)，儿童阅读与学习研究院，北京 100875) (2 深圳市红岭小学，深圳 518000) (3 河南师范大学教育学部，新乡 453007)

摘 要 两两比较作为一种相对整体并能够高效获得可靠结果的文本难度评估手段，在汉语文本评估中的效果还有待探索。本研究通过80名评估者对80篇文本的两两比较，探讨了比较次数对两两比较信效度的影响。结果表明，两两比较获得的结果具有较高的信度，且与文本册数和可读性分数显著相关，随着比较次数的增加，信效度逐渐增加且趋于稳定，同时未发现评估者特征对比较结果的影响，这意味着两两比较在汉语文本难度评估中具备一定可靠性。

关键词 汉语分级阅读，文本难度，两两比较，小学语文教材。

分类号 B842

1 引言

阅读是日常生活与个人发展的必备技能。自2014年起，全民阅读这一理念连续多次被写入政府工作报告之中，充分显示了国家对推广阅读、提升国民文化素养的重视与决心。然而，2021年我国未成年人的年均阅读量仅为10.93本书(中国新闻出版研究院, 2022)，这一数字显著低于美国年均阅读量的28.4本(Renaissance, 2022)。这一现象背后，一个不可忽视的原因或许是缺乏科学的中文分级阅读体系。面对规模庞大的出版物市场，读者往往难以精准筛选出适合自己阅读水平的优质书籍。

1.1 分级阅读与文本难度

小学阶段是培养儿童阅读能力的关键时期(Kuhn et al., 2010)，在这一时期为儿童选择适宜的文本至关重要。若文本太简单，儿童阅读能力难以得到训练和提高；若文本太难，则可能导致儿童受挫，损害其阅读动机(Chall & Conard, 1991)。因此，需要将阅读材料按照难度梯度进行组织，根据读者水平匹配合适的文本并给予针对性指导，即进行分级阅读(Fountas & Pinnell, 2012)，而构建分级阅读体系的难点之一在于评估文本难度。本研究致力于探索两两比较在汉语文本难度评估中的信度和有效性，从而为汉语分级阅读体系的构建提供新的文本难度评估思路。

文本难度指的是文本易于阅读和理解的难易程度(Dale & Chall, 1949)。Coertjens等人(2017)从两个维度对文本难度的评估方法进行了分类：绝对或相对(absolute vs. comparative)，整体或分解(holistic vs. analytic)。其中，绝对或相对指的是文本评估的结果，前者是对单个文本评分，得到的是难度量尺上的一个数值，后者则是比较两个或以上文本，得到的是二者在难度量尺上的先后顺序；整体或分解指的是文本评估的性质，前者是评估整篇文本的总体情况，后者评估文本多个方面的特征。由此产生了四类文本评估方法(见图1)：绝对整体评估、相对整体评估、绝对分解评估、相对分解评估。

常见的文本难度评估方法包括专家评估和可读性公式。其中，专家评估属于绝对整体评估，依赖专家团队对文本的难度进行主观评级。目前，教材中的每一篇课文对应的适读年级，即教材册数，就是专家评估的结果，直接反映了课文内容的难度层次与阅读适宜性。而可读性公式属于绝对分解评估，主要基于文本字、词、句、篇层面的特征，通过数学模型预测文本的可读性分数，从而标定文本的难度。广为人知的蓝思分级阅读框架(the Lexile Framework; Smith et al., 1989)就是以词频和句长两个文本特征建立文本可读性公

收稿日期：2025-03-18

* 基金项目：国家语委“十四五”科研规划项目(WT45-41)；国家社会科学基金教育学一般项目“AI赋能因材施教：基于智能体的学习困难的多层级评估与个体化干预体系研究”(BBA250057)。

通讯作者：李 虹，E-mail：psy.lihong@bnu.edu.cn。

式，以此评估文本难度。

尽管专家评估和可读性公式已经被广泛运用并取得了不错的效果，但仍存在不足之处。专家评估过于依赖评估者的经验，使得不同专家评估时的内部标准可能不同，存在评估者差异，且难以进行大规模的语料库建设和标定，也难以进行量化分析 (Bloxham, 2009; Fry, 2002)。而可读性公式多基于文本中较为浅层的特征 (陈茹玲 等, 2015; 刘苗苗 等, 2021)，容易忽视更深层次的因素，如文本立意等，导致不能真实反映读者的心理现实。此外，可读性公式的开发和验证需要基于特定的有难度指标的黄金语料库，这也使得可读性公式的外部效度难以验证，无法推广到更广泛的一般性阅读材料中 (Sheehan et al., 2014)。而两两比较作为一种相对的、整体的评估方法，可以有效弥补专家评估和可读性公式的局限。

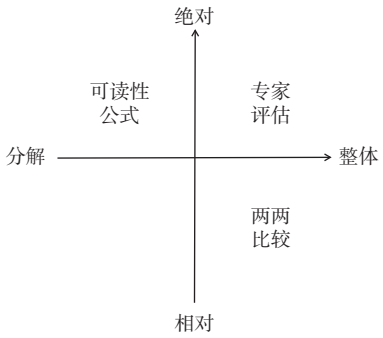

图 1 文本评估方法分类

两两比较有两个关键假设 (Crompvoets et al., 2020; Paquot et al., 2022)：(1) 人们对两个对象做相对判断时，要比单独给一个对象评分更简单而高效，也更可信；(2) 两两比较时，评估者依靠对对象的整体认知做出判断。对于假设 1，研究者发现评估者比较两篇文本时，比单独评估一篇文本时更轻松，评估结果也更为可靠 (Pollitt, 2012)，并且相较于专家评估，两两比较不会过分依赖于评估者的专业背景，就像“不怕不识货，就怕货比货”一样，普通人也能做出客观可靠的判断，从而可以实现高效评估大量文本的目的。对于假设 2，两两比较的整体性有助于度量较为复杂的属性，而文本难度就是一种复杂属性。可读性公式中词频、句长等表面特征的集合可能无法完整地表征文本难度，而文本更深层次的特征 (如作者情感、文章主旨) 又难以直接测量，两两比较的整体性有助于更加全面地评估文本难度，真实反映读者的心理感受。同时，两两比较的结果是综合所有评估者对文本的判断得到的文本难度评定等级，代表了评估者们对文本难度的共识 (Jones & Inglis, 2015)，能够减少个体差异带来的影响，具有良好的实际应用前景。

1.2 两两比较

两两比较最初由 Thurstone(1927) 提出，最初主要是在决策分析、排名、偏好建模等领域中使用。该方法要求评估者比较两个对象，判断哪一个对象在某个属性上表现得更好。当得到了海量的两两比较数据后，可以通过 Bradley-Terry-Luce (BTL) 模型 (Bradley & Terry, 1952; Luce, 1959)，预测两两比较中一个对象优于另一个对象的概率，根据概率大小可以对对象进行排序，从而获得所有比较对象在某一属性上的评定等级，即两两比较分数。

Pollitt 和 Murray(1996) 首次将两两比较的方法应用于教育领域，用于评估学生口语水平。随后，Pollitt(2012) 进一步将其应用到英文文本评估中，如评估学生作文质量 (Bramley & Vitello, 2019; Lesterhuis et al., 2018; Pollitt, 2012; Thwaites et al., 2024; Wheadon et al., 2020)、试卷的相对难度 (Jones et al., 2015) 等，近几年有研究者开始尝试将两两比较用于英文文本难度的评估中 (Crossley et al., 2023; Crossley et al., 2019)，但将这种方法应用于汉语文本的难度评估仍有两个问题有待探索。

其一是文本比较次数对两两比较结果可靠性的影响。一方面，更多信息能使 BTL 模型中的估计值更加准确可靠，评估者评估的次数越多，对比较任务越熟悉，也能有效提升结果的可靠性。另一方面，过多的比较次数可能会让评估者感到疲劳，进而影响评估结果的有效性。就信度而言，Bramley(2007) 和 Pollitt(2012) 认为，在两两比较中存在一个最小阈值，只要比较次数达到该阈值，就能获得足够多的信息来估算每个比较对象的两两比较分数。模拟研究分别对 20、25、30 和 100 篇文本进行两两比较，结果发现，平均每篇文本需要至少进行 20 次比较才能达到 0.80 的可靠性 (Crompvoets et al., 2020)。Verhavert 等人 (2019) 的元分析显示，评估对象被比较 10 ~ 14 次才能达到 0.70 的信度，比较 26 ~ 37 次才能达到 0.90 的信度。由此可见，两两比较结果的信度会受到比较次数的影响。而在效度方面，尽管有研究已经证实了在评估论文写作质量时，两两比较的结果与教师评分 (包括绝对整体评估、绝对分解评估) 的

结果存在中等强度的相关 (Landrieu et al., 2022; Thwaites et al., 2024)，但目前尚未有研究探索比较次数对两两比较结果效度的影响。系统探究文本比较次数与信度 (评估结果的一致性和稳定性) 和效度 (评估结果的准确性和真实性) 的关系，才能找到效率–可靠性的权衡点。

其二是采用两两比较评估文本难度时，评估者的相关知识以及评估经验是否会影响评估结果。针对这一问题，已有研究并未得出完全一致的结论。一些研究发现，评估者的相关知识和评估经验对两两比较的决策依据 (Lesterhuis et al., 2022) 和结果信度 (Verhavert et al., 2019) 无影响。另一些研究则发现评估者的文化背景会对结果产生影响 (Bartholomew et al., 2020)。Paquot 等人 (2022) 则通过引入拟合统计量 infit 值来衡量评估者个体判断与群体共识之间的一致程度，并通过比较不同评估者群体的 infit 值发现，是否接受过正式评估培训会对两两比较产生影响。然而，这些不一致的结果可能并不仅仅是因为评估者特征之间的差异，而可能源于研究设计的不同。Verhavert 等人的元分析结果表明，专家评估者和新手评估者的评估结果之间不存在显著差异，评估者的专业水平只会影响获得可靠结果所需要的比较次数，专家评估者只需约 26 次，新手评估者则需要约 37 次。而在 Paquot 等人的研究中，每个评估对象仅被比较 16 ~ 22 次，有限的比较次数可能使得新手评估的结果尚未达到稳定，从而显示出与专家评估结果的差异。此外，评估者的同质性也可能影响分析结果。Lesterhuis 等人的研究中，招募的专业评估者均为教师或考官，而 Bartholomew 等人的研究中，评估者为来自美国、英国和瑞典三个不同国家的设计行业从业者，评估者文化背景差异可能是导致评估者特征影响两两比较结果的原因之一。

为了回答两两比较是否可用于汉语文本难度评估这一核心问题，本研究将邀请不同背景的大学生，完成汉语文本的两两比较任务，把不同比较次数的结果与文本册数、文本可读性分数相比较，探究以下两个问题：(1) 文本至少被比较多少次才能获得可信而有效的两两比较结果？(2) 两两比较结果是否受到评估者特征 (专业、高考语文成绩和相对阅读习惯) 的影响？

2　研究方法

2.1　被试

本研究的评估者为 80 名在校大学生，均为汉语母语者，平均年龄为 20.09 岁 (*SD*=2.02)，其中男生 5 人，女生 75 人。所有人均自我报告了高考语文成绩和专业类别并填写了相对阅读习惯问卷。

2.2　研究材料

2.2.1　相对阅读习惯问卷

相对阅读习惯问卷 (Comparative Reading Habits; Chen & Fang, 2015) 由 5 道题目构成，要求个体对自己相对于同龄人在阅读时长、阅读材料复杂度、阅读喜爱度、阅读速度和阅读理解度 5 个方面的表现进行 7 点李克特评分，1 分代表“完全不符合”，7 分代表“完全符合”。不同组别评估者的相对阅读习惯分数如表 1 所示。

表 1　评估者相对阅读习惯分数、infit 值描述统计 [*M*(*SD*)] 及分组检验结果

分组类别	分组依据	人数	相对阅读习惯分数	infit值	*t*/*Z*	*p*
专业	理科	29	21.45(3.26)	0.90(0.31)	–0.31	0.76
	文科	51	21.31(4.94)	0.93(0.40)		
高考语文成绩	119分及以下	38	21.29(5.29)	0.98(0.41)	–1.40	0.16
	120分及以上	42	21.43(3.44)	0.87(0.32)		
相对于阅读习惯	≤21	35	17.54(2.95)	0.96(0.38)	–0.94	0.35
	>21	45	24.33(2.67)	0.89(0.36)		
总计		80	21.36(4.38)	0.92(0.37)		

2.2.2　两两比较文本

从小学 1 ~ 4 年级的语文教材 (共 8 册) 中选取 80 篇课文 (其中 18 篇选自北师大版、23 篇选自人教版、16 篇选自苏教版、23 篇选自西南师大版语文教材)，每册各 10 篇，不包含古诗、古文和现代诗。文本信息如表 2 所示，文本字数 (文本中所有汉字的个数)、字种 (文本中所有汉字的种类数，即重复出现的汉字只计一次)、词数 (文本中所有词汇的个数)、词种 (文本中所有词汇的种类数，即重复出现的词汇只计一次)、句数 (文本中以句号、问

号、感叹号、破折号和省略号结尾的句子数)、句长 (文本总字数/完整句子数) 均随着册数的增长而逐渐增长。此外，为每一篇文本编写一道事实性判断的阅读理解题目，用于促使评估者认真阅读文本。

为了衡量两两比较评估结果的准确性，参照 Liu 等人 (2024) 的可读性公式，计算了每篇文本的可读性分数作为效标。具体而言，可读性分数是以小学语文教材的册数为难度指标，通过对每篇课文提取 265 个文本特征，采用随机森林和支持向量机两种机器学习算法所构建的可读性模型估算得到 (Liu et al., 2024)，分数越大表示文本越难。

表 2　两两比较文本基本信息 [*M*(*SD*)]

册数	文本篇数	字数	字种	词数	词种	句数	句长	可读性分数	
								随机森林	支持向量机
1	10	70.20(36.67)	35.30(16.28)	52.40(29.27)	29.10(14.42)	4.30(2.00)	18.03(8.66)	1.24(0.13)	1.67(0.27)
2	10	115.90(55.67)	60.00(14.71)	87.30(44.49)	50.60(14.28)	7.20(4.21)	17.56(4.55)	2.04(0.22)	2.30(0.17)
3	10	163.60(48.80)	88.50(16.35)	118.30(39.34)	73.80(16.35)	9.40(3.60)	19.03(6.80)	2.97(0.17)	3.33(0.24)
4	10	237.10(61.03)	106.10(23.85)	174.80(45.47)	91.20(21.52)	13.90(5.26)	18.04(3.98)	3.86(0.15)	4.18(0.13)
5	10	361.90(114.09)	158.50(36.24)	253.10(87.20)	138.40(39.39)	19.40(7.63)	19.46(3.42)	5.06(0.25)	5.11(0.10)
6	10	434.40(54.73)	199.40(21.19)	300.40(45.40)	169.50(22.85)	17.30(5.33)	26.56(6.12)	6.22(0.16)	6.20(0.17)
7	10	513.20(66.14)	241.30(17.52)	333.80(51.20)	203.50(21.68)	21.50(4.77)	24.48(3.62)	7.23(0.15)	7.25(0.10)
8	10	716.10(175.72)	303.80(37.35)	458.40(127.33)	256.90(44.58)	24.60(8.07)	30.05(5.06)	8.36(0.24)	8.00(0.25)
总计	80	326.55(224.95)	149.11(91.24)	222.31(145.33)	126.63(78.80)	14.70(8.56)	21.65(6.92)	4.62(2.39)	4.75(2.17)

2.3 两两比较任务

两两比较实验任务中，屏幕上每次同时呈现两篇文本和三个 2 选 1 的问题，如图 2 所示。其中，前 2 个问题分别是针对文本 1 和文本 2 的阅读理解题，第 3 个问题为“哪篇文本读起来更难？”，评估者按照自己的理解进行选择，程序自动记录选择结果，并在评估者点击“下一个”时进入下一个试次。

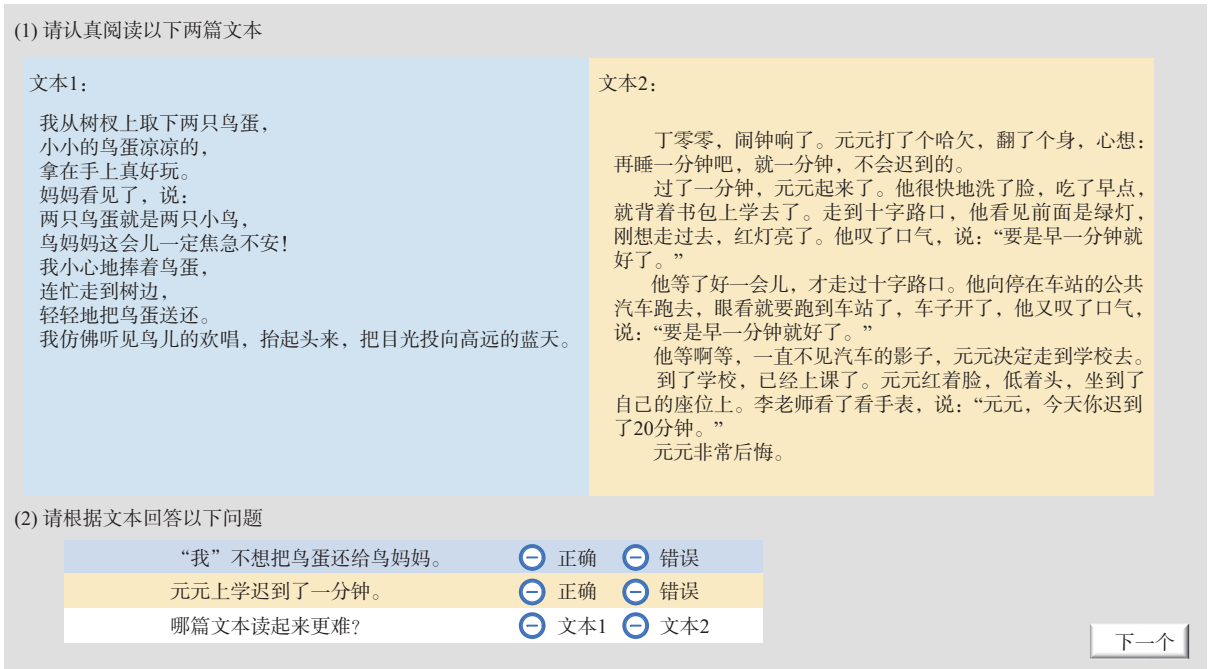

图 2　两两比较任务界面

注：示意图中左侧文本 (文本 1) 为人教版小学语文第二册 (2001 年版) 的课文，右侧文本 (文本 2) 为人教版小学语文第三册 (2001 年版) 的课文。

文本呈现顺序采用 Paquot 等人 (2022) 研发的自适应两两比较 (Adaptive Comparative Judgment, ACJ) 程序，其基本原则是：(1) 选择所有 80 篇文本中比较次数最少的一篇文本作为文本 1；(2) 寻

找已有两两比较结果中与文本 1 比较时获胜概率最接近 50%，且比较次数最少的一篇文本作为文本 2；(3) 避免同一名评估者再次对同样的两篇文本进行两两比较。最终，每名评估者比较的文本对取决于已完成比较的结果，每篇文本的比较对象随着比较轮次的迭代，也更加集中在与之难度差异较小的文本之中。这样既能避免读者耗费精力在难度差异较大的文本间，又能使得难度接近的文本得到多位评估者的反复比较，从而提高评估的效率和代表性。最终，每名评估者均完成 40 次两两比较，每篇文本被比较了 79 ~ 81 次 (M=80.00, SD=0.42)。

2.4　研究流程

首先，通过问卷收集评估者的人口学信息、高考语文成绩、专业、相对阅读习惯等与个体汉语阅读能力有关的信息，约 5 分钟。随后，评估者先进行 3 次两两比较练习，以保证其理解了任务要求。正式实验中，每位评估者共需完成 40 次比较，大约需要 60 分钟，中途休息 3 分钟。

2.5　数据分析

首先，参考前人方法 (Verhavert et al., 2019; Verhavert et al., 2018)，以两两比较的尺度分离信度 (scale separation reliability, SSR) 作为信度指标。其次，以文本册数、文本可读性分数作为效标，通过方差分析和相关分析，探索两两比较的效度。最后，对 BTL 模型的拟合度参数 infit 值进行差异检验，以此分析评估者特征对两两比较结果的影响。

3　结果

3.1　两两比较的信度

SSR 是 Rasch 模型中测量信度的方法 (Bramley, 2015)，代表了个体与总体的偏离程度。元分析发现 SSR 在两两比较中能很好地反映评估者的内部一致性 (Verhavert et al., 2018) 和文本两两比较分数的稳定性 (Pollitt, 2012)，其数值越大，代表一致性程度越高。为了评估两两比较的信度及其与文本被比较次数之间的关系，本研究计算了本次实验的 SSR 值及其随着文本被比较次数增加而变化的趋势，如图 3 所示。

结果显示，随着文本被比较次数的增多，SSR 逐渐增大并趋于稳定。当文本仅被比较 17 次时，SSR 已经高于 0.90，并且在此后连续三次比较中变化量保持在 0.01 以下，这说明 17 次的两两比较就已经具有良好且稳定的信度。当比较次数达到 80 次时，SSR 达到 0.98。

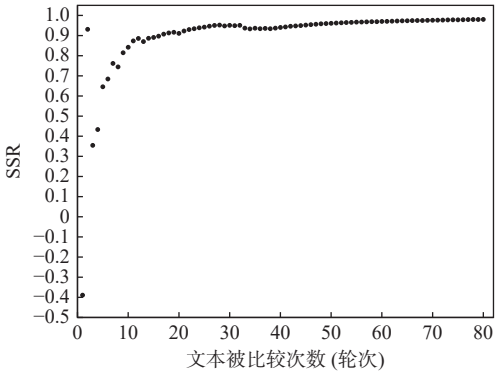

图 3　SSR 随着文本被比较次数的变化

3.2　两两比较的效度

为了探究文本比较次数与两两比较效度的关系，统计了文本两两比较分数随着文本被比较次数增加而变化的趋势。本研究中两两比较分数范围为 [–6, 6]，代表了该文本在 80 篇文本中的相对位置。两两比较分数越大，代表文本难度越低，越容易理解；分数越小，代表越难理解。

80 篇文本两两比较分数的描述统计如表 3 所示，随着文本册数的增高，文本两两比较分数逐渐降低，表明文本难度升高。为了比较不同册数间的分数差异，以册数为自变量，文本两两比较分数为因变量，分别对不同比较次数下的结果进行了单因素方差分析。结果表明，不同册数之间的文本两两比较分数差异显著。具体结果如下，文本被比较 20 次时，$F(7, 72)=10.21$，$p<0.001$；40 次时，$F(7, 72)=19.53$，$p<0.001$；60 次时，$F(7, 72)=26.35$，$p<0.001$；80 次时，$F(7, 72)=44.86$，$p<0.001$。进一步分析发现，随着文本被比较次数的增多，相邻册数间文本两两比较分数的差异逐渐明显，分数变化的趋势更加明显而稳定，表明更多的比较次数可以增加文本难度评估的有效性。

表 3　不同比较次数下的两两比较分数描述统计 [*M*(*SD*)]

册数	文本篇数	两两比较分数			
		R=20	R=40	R=60	R=80
1	10	2.99^{a}(0.87)	2.88^{a}(1.03)	3.27^{a}(1.11)	3.66^{a}(1.03)
2	10	0.85^{b}(1.19)	0.67^{b}(1.06)	1.50^{b}(1.47)	2.19^{ab}(1.08)
3	10	0.65^{bc}(1.68)	0.48^{bc}(1.13)	1.27^{b}(0.91)	1.48^{b}(1.05)
4	10	-0.73^{bc}(1.50)	-0.08^{bc}(1.13)	-0.05^{bc}(1.32)	-0.18^{c}(1.01)
5	10	-0.59^{bc}(1.39)	0.24^{bc}(1.17)	-0.42^{c}(1.11)	-0.76^{cd}(1.07)
6	10	-0.40^{bc}(1.39)	-0.92^{cde}(1.06)	-1.26^{cd}(0.88)	-1.17^{cd}(0.99)
7	10	-1.67^{c}(1.41)	-1.29^{de}(0.57)	-1.65^{cd}(1.11)	-1.80^{d}(1.41)
8	10	-1.10^{bc}(1.95)	-1.99^{e}(1.24)	-2.67^{d}(1.46)	-3.41^{e}(1.00)

注：每一列数据相同上标表示差异不显著，不同上标表示差异显著；R代表文本被比较的次数(轮次)，以下同。

为了进一步对比两两比较分数与文本册数的关系及比较次数对两两比较效度的影响，分别以文本被比较 20 次、40 次、60 次、80 次的两两比较分数为纵坐标，以文本册数为横坐标，绘制散点图，如图 4 所示。

图 4a、图 4b、图 4c、图 4d 分别代表了文本被比较 20 次、40 次、60 次和 80 次时文本两两比较分数与册数的分布，每一个点代表一篇文本。由此可见，随着文本被比较次数的增加，文本分数按册分布递减的规律更加清晰。

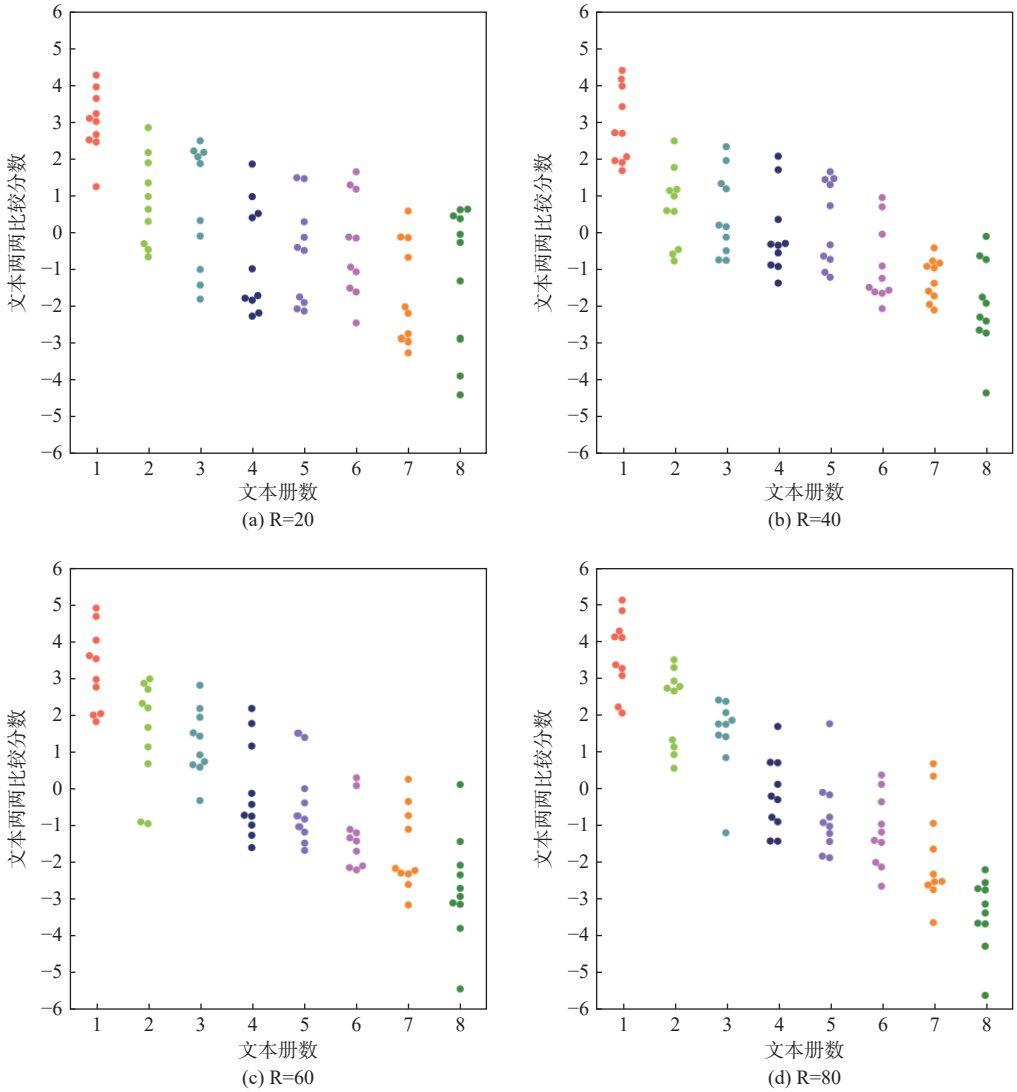

图 4　两两比较分数与册数的散点图

进一步将文本按照两两比较分数由高到低分为 8 个类别，每个类别包含 10 篇文本。将该类别与册数进行对比，绘制混淆矩阵热图，如图 5 所示。图 5a、图 5b、图 5c、图 5d 分别代表了文本被比较 20 次、40 次、60 次和 80 次时类别与册数的分布，方块中的数字代表处于该类别等级的文本数量。

从图 5 的变化可以看出，随着文本被比较次数的增加，文本类别与册数的对应关系逐渐收敛，趋于一致，文本更加集中地分布在对角线附近。

最后，分别计算了文本被比较 20 次、40 次、60 次和 80 次时，两两比较分数与可读性分数、册数的相关系数。结果如表 4 所示，文本被比较 20 次、40 次、60 次和 80 次时，文本两两比较分数与可读性分数、册数的相关均显著 (p<0.01)。对相关系数进行 Fisher r-to-z 转换 (Meng et al., 1992)，对相关系数之间的差异进行单尾检验，结果显示随着比较次数的增加，文本两两比较分数与可读性分数、册数的相关系数都有明显提升，比较次数越多，相关越高，两两比较分数与随机森林可

读性分数的相关在比较 60 次时，与比较 80 次时的相关系数无显著差异 (Z=1.49, p=0.07)。综上所述，在文本被比较 60 次时，文本两两比较分数具有较高且稳定的信度和效度。

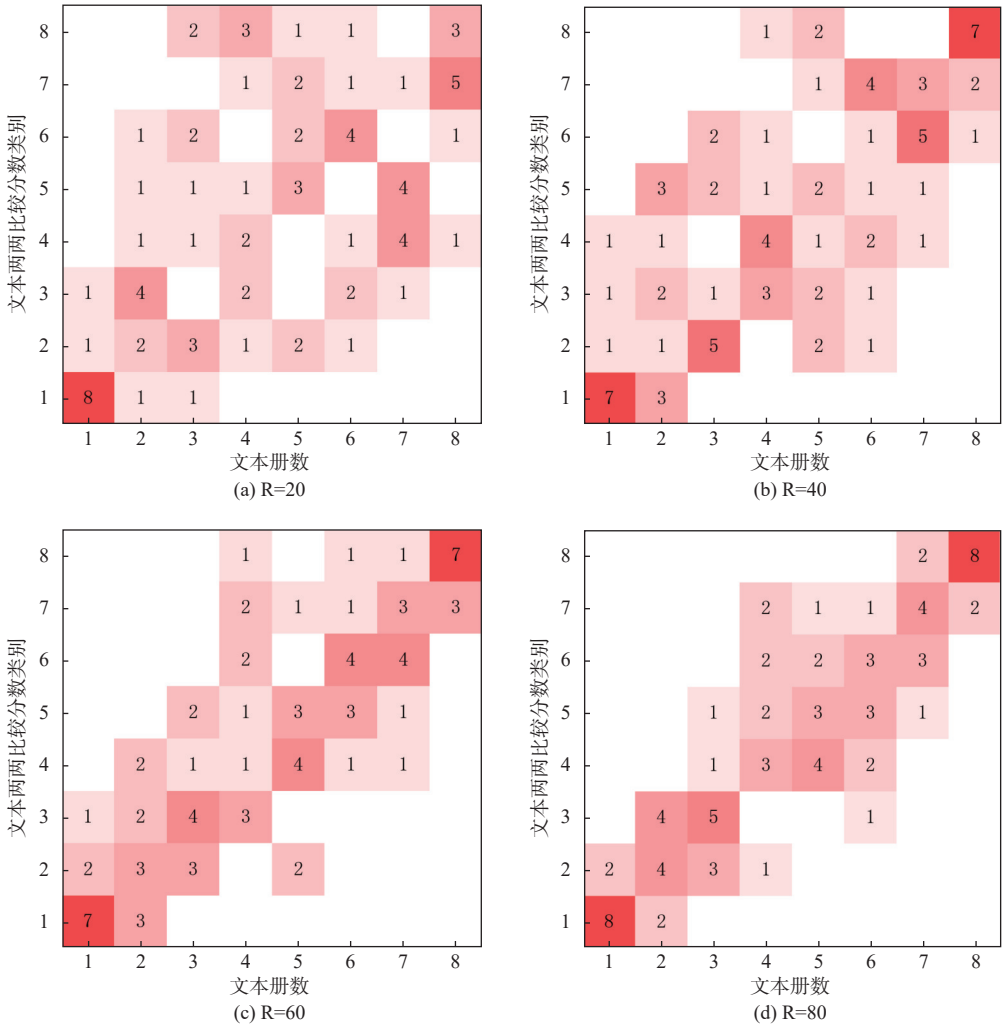

图 5　两两比较分数类别与册数的混淆矩阵热图

表 4　不同文本难度指标的相关 (n=80)

	1	2	3	4	5	6
1.文本两两比较分数(R=20)						
2.文本两两比较分数(R=40)	0.52**					
3.文本两两比较分数(R=60)	0.55**	0.73**				
4.文本两两比较分数(R=80)	0.57**	0.68**	0.83**			
5.文本册数	–0.59**	–0.76**	–0.84**	–0.89**		
6.文本可读性分数(随机森林)	–0.58**	–0.78**	–0.86**	–0.90**	0.99**	
7.文本可读性分数(支持向量机)	–0.58**	–0.78**	–0.83**	–0.89**	0.99**	0.99**

注：**p<0.01。

3.3　评估者特征对两两比较结果的影响

为了探究评估者特征对两两比较结果的影响，本研究对不同专业组、高考语文成绩高低组和相对阅读习惯高低组评估者的 infit 平均值进行了差异检验，描述统计见表 1。infit 值代表了个体与 BTL 模型的拟合度。infit=1 时，代表个体与模型完全拟合 (杨慷 等, 2021)，infit 值偏离总体均值两个标准差以外的个体被视为拟合程度较差 (Pollitt,

2012)。在本研究中，一共发现 4 位拟合程度较差的评估者，但删除 4 人的数据后，SSR 与文本比较次数的关系、读者特征差异等结果均与删除前一致。根据报告所有数据的原则，所有分析中仍然包含了这 4 人的数据。

如表 1 所示，由于不同高考语文成绩和不同相对阅读习惯的评估者的 infit 值分布不服从正态分布，对二者采用 Mann-Whitney U 检验，对不同专业组的 infit 值做独立样本 t 检验。结果表明，不同专业组评估者的 infit 值之间不存在显著差异 (t=–0.31, p=0.76)，高考语文成绩高低组评估者的 infit 值之间不存在显著差异 (Z=–1.40, p=0.16)；相对阅读习惯得分高低组评估者的 infit 值之间也不存在显著差异 (Z=–0.94, p=0.35)。

综合上述结果可知，本研究未能找到评估者的专业、高考语文成绩和相对阅读习惯会影响两两比较结果的证据。

4 讨论

4.1 两两比较用于汉语文本难度评估的信效度

本研究旨在探索两两比较在汉语文本难度评估中的可行性。就信度而言，本研究分析发现 SSR 在文本被比较 17 次时便达到 0.90 以上且趋于稳定，这一结果优于前人元分析 (Verhavert et al., 2019) 结论 (比较次数在 26 ~ 37 之间才能达到 0.90 的信度)，也优于 Crossley 等人 (2023) 的英语文本难度评估中的结果 (两两比较 46 次之后，评估结果的分半信度为 0.85)。文本材料和比较方法的选择，可能是造成本研究中 SSR 达到稳定时的文本被比较次数更少，且信度更高的原因。就文本材料而言，本研究的实验材料来自小学 1 ~ 4 年级的课本，理论上文本本身就具有一定难度差异，低年级课文简单，而高年级课文复杂，使得评估者更容易做出判断，而 Crossley 等人的实验材料来自开放图书库里的文本选段，本身的难度差异可能没有课文那样直观而明显。就比较方法而言，本研究采用的是自适应的两两比较，而元分析中的 49 个比较结果均未使用自适应算法。Paquot 等人 (2022) 采用自适应两两比较评估学生作文写作质量时，发现比较 14 次时 SSR 即可达到 0.95 的水平，这与本研究结果相一致，意味着自适应算法减少了难度差异过大的文本间的比较次数，提高了两两比较的效率，值得今后的研究者借鉴采用。此外，本研究中，文本被比较 80 次后的 SSR 高达 0.98，这可能意味着文本被比较次数过多。这也提示研究者，可以通过实时监测每一轮比较结束后的 SSR，当 SSR 趋于稳定时，可以考虑停止实验，最大化地提高效率。

在效度方面，本研究发现随着文本比较次数的增加，两两比较结果的效度逐渐提高。在本研究中，文本被比较 20 次时，两两比较分数与文本册数 (绝对整体评估)、可读性分数 (绝对分解评估) 就具有显著相关，这与前人研究 (Thwaites et al., 2024) 在作文评估中的发现一致，两两比较进行约 25 次后就与专家评估结果达到中等强度相关 (r=0.66)。然而以往的研究 (Landrieu et al., 2022; Thwaites et al., 2024) 未能进一步探究当比较次数增多后，两两比较的效度如何变化。Crossley 等人 (2023) 在英文文本难度评估中发现，文本被比较约 46 次时，两两比较结果与可读性公式结果具有中等强度相关 (r=0.56)。而本研究发现，文本被比较 60 次时，上述相关系数达到了 0.83，并且与比较 80 次时的相关系数没有显著差异，这可能意味着，尽管文本被比较 17 次后，两两比较的信度达到稳定，但要进一步提高效度，还可以增加文本被比较的次数。根据本研究的结果，本文比较 60 次时能获得较为可信且有效的两两比较结果。

此外，尽管两两比较与专家评估、可读性公式的结果存在中高等强度的相关，但不同评估方法间的结果也存在差异，如两两比较中难度最低的 10 篇文本并不全是第一册的 10 篇文本，这可能意味着针对文本难度，不同的评估方法侧重点不同 (Landrieu et al., 2022; Thwaites et al., 2024)，也可能反映了两两比较能从整体上更加全面地评估文本这一基本假设。Lesterhuis 等人 (2022) 和 Lesterhuis 等人 (2018) 的研究表明，评估者在进行两两比较判断议论文写作质量时，相较于拼写或语法错误等表面因素，会更多地考虑文本组织、论证过程等深层因素。对于文本难度这一综合性、主观性很强的概念而言，两两比较可作为评估方法之一。

4.2 评估者特征对两两比较的影响

本研究未找到评估者的专业、高考语文成绩和相对阅读习惯对两两比较结果存在显著影响的证据，这基本符合前人研究的结论 (Lesterhuis et al., 2022; Verhavert et al., 2019)，即使是没有接受过专业培训的评估者也能做出可靠的两两比较判断，再次证实两两比较对评估者的专业性要求不高，提高了文本评估的效率。此外，两两比较的结果

代表了评估者们的共识，削弱了主观因素的干扰，这也使得两两比较更加适用于评估无法给出具体评分标准的复杂构念，如文本难度。

5　结论

本研究采用两两比较对小学 1 ~ 4 年级语文教材中的 80 篇文本进行难度评估，结果表明两两比较的结果具有较高的信度，且与文本册数和可读性分数具有显著相关，同时，未发现评估者特征对比较结果的影响。综合而言，两两比较在汉语文本评估中具有一定可靠性、有效性和较为广泛的适用性，可以作为汉语文本分级阅读研究的新方法。

参 考 文 献

陈茹玲, 蔡鑫廷, 宋曜廷, 李宜宪. (2015). 文本适读性分级架构之建立研究. *教育科学研究期刊*, *60*(1), 1–32.

刘苗苗, 李燕, 王欣萌, 甘琳琳, 李虹. (2021). 分级阅读初探: 基于小学教材的汉语可读性公式研究. *语言文字应用*, (2), 116–126.

杨慷, 贺文洁, 王海龙. (2021). 单参数单维度 Rasch 模型的优势与意义. *心理科学*, *44*(6), 1491–1498.

中国新闻出版研究院. (2022). *第十九次全国国民阅读调查结果*. 2022-11-30 取自 https://society.huanqiu.com/article/47ix20UIt5x

Bartholomew, S. R., Ruesch, E. Y., Hartell, E., & Strimel, G. J. (2020). Identifying design values across countries through adaptive comparative judgment. *International Journal of Technology and Design Education*, *30*(2), 321–347.

Bloxham, S. (2009). Marking and moderation in the UK: False assumptions and wasted resources. *Assessment & Evaluation in Higher Education*, *34*(2), 209–220.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3–4), 324–345.

Bramley, T. (2007). Paired comparison methods. In P. Newton, J. A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 246–300). London: Qualifications and Curriculum Authority.

Bramley, T. (2015). *Investigating the reliability of adaptive comparative judgment*. Cambridge: Cambridge University Press & Assessment.

Bramley, T., & Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *26*(1), 43–58.

Chall, J. S., & Conard, S. S. (1991). *Should textbooks challenge students? : The case for easier or harder books*. New York: Teachers College Press.

Chen, S. Y., & Fang, S. P. (2015). Developing a Chinese version of an author recognition test for college students in Taiwan. *Journal of Research in Reading*, *38*(4), 344–360.

Coertjens, L., Lesterhuis, M., Verhavert, S., van Gasse, R., & De Maeyer, S. (2017). Judging texts with rubrics and comparative judgement: Taking into account reliability and time investment. *Pedagogische Studien*, *94*(4), 283–303.

Crompvoets, E. A. V., Béguin, A. A., & Sijtsma, K. (2020). Adaptive pairwise comparison for educational measurement. *Journal of Educational and Behavioral Statistics*, *45*(3), 316–338.

Crossley, S., Heintz, A., Choi, J. S., Batchelor, J., Karimi, M., & Malatinszky, A. (2023). A large-scaled corpus for assessing text readability. *Behavior Research Methods*, *55*(2), 491–507.

Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, *42*(3–4), 541–561.

Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English*, *26*(1), 19–26.

Fountas, I. C., & Pinnell, G. S. (2012). Guided reading: The romance and the reality. *The Reading Teacher*, *66*(4), 268–284.

Fry, E. (2002). Readability versus leveling. *The Reading Teacher*, *56*(3), 286–291.

Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: Can comparative judgement help? *Educational Studies in Mathematics*, *89*(3), 337–355.

Jones, I., Swan, M., & Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education*, *13*(1), 151–177.

Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly*, *45*(2), 230–251.

Landrieu, Y., De Smedt, F., van Keer, H., & De Wever, B. (2022). Assessing the quality of argumentative texts: Examining the general agreement between different rating procedures and exploring inferences of (dis)agreement cases. *Frontiers in Education*, *7*, 784261.

Lesterhuis, M., Bouwer, R., van Daal, T., Donche, V., & De Maeyer, S. (2022). Validity of comparative judgment scores: How assessors evaluate aspects of text quality when comparing argumentative texts. *Frontiers in Education*, *7*, 823895.

Lesterhuis, M., van Daal, T., van Gasse, R., Coertjens, L., Donche, V., & De Maeyer, S. (2018). When teachers compare argumentative texts: Decisions informed by multiple complex aspects of text quality. *L1-Educational Studies in Language and Literature*, *18*(1), 1–22.

Liu, M. M., Li, Y. X., Su, Y. Q., & Li, H. (2024). Text complexity of

Chinese elementary school textbooks: Analysis of text linguistic features using machine learning algorithms. *Scientific Studies of Reading*, *28*(3), 235–255.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: John Wiley & Sons, Inc.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*(1), 172–175.

Paquot, M., Rubin, R., & Vandeweerd, N. (2022). Crowdsourced adaptive comparative judgment: A community-based solution for proficiency rating. *Language Learning*, *72*(3), 853–885.

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *19*(3), 281–300.

Pollitt, A., & Murray, N. L. (1996). *What raters really pay attention to*. In M. Milanovic & N. Saville (Eds.), *Studies in language testing 3: Performance testing, cognition and assessment* (pp. 74–91). Cambridge: Cambridge University Press.

Renaissance. (2022). *What kids are reading report 2022*. Retrieved November 30, 2022, from https://www.renaissance.com/2022/03/01/news-renaissance-shares-findings-of-worlds-largest-annual-k12-reading-survey/

Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The TextEvaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal*, *115*(2), 184–209.

Smith, D. R., Stenner, A. J., Horabin, I., & Smith, M. (1989). *The lexile scale in theory and practice: Final report for NIH Grant HD-19448*. Bethesda, MD: National Institutes of Health.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286.

Thwaites, P., Kollias, C., & Paquot, M. (2024). Is CJ a valid, reliable form of L2 writing assessment when texts are long, homogeneous in proficiency, and feature heterogeneous prompts? *Assessing Writing*, *60*, 100843.

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *26*(5), 541–562.

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2018). Scale separation reliability: What does it mean in the context of comparative judgment? *Applied Psychological Measurement*, *42*(6), 428–445.

Wheadon, C., Barmby, P., Christodoulou, D., & Henderson, B. (2020). A comparative judgement approach to the large-scale assessment of primary writing in England. *Assessment in Education: Principles, Policy & Practice*, *27*(1), 46–64.

The Application of the Comparative Judgment in Chinese Text Difficulty Assessment

YANG Guandoudou [1], TAN Jingwen [1,2], LIU Miaomiao [1,3], LI Hong [1]

(1 Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education(Beijing Normal University), Institute of Children's Reading and Learning, Faculty of Psychology, Beijing Normal University, Beijing 100875; 2 Shenzhen Hongling Primary School, Shenzhen 518000; 3 Faculty of Education, Henan Normal University, Xinxiang 453007)

Abstract

One effective way to obtain a trustworthy assessment of text difficulty is to employ comparative judgment, a holistic assessment method that has not yet been utilized in Chinese research. The current study gathered 80 evaluators and evaluated the difficulty of 80 texts using the comparative judgment approach. The results showed that the results obtained from comparative judgment had high reliability and were significantly correlated with the number of volumes and readability scores. The reliability and validity tended to be stable with the increase in the number of comparative judgments. No effect of the evaluator characteristics on the comparative judgment results was found, which implies that comparative judgment is also reliable in the assessment of the difficulty of Chinese texts.

Key words Chinese leveled reading, text difficulty, comparative judgment, elementary-school-level Chinese language and literature textbooks.