互联网教育研究院
China Online Education Research Institute

口语 100 产品测评报告

http://www.medu.org.cn/

微信：medu-org-cn

北京朝阳德外大街华严里甲 1 号

一、测评背景

智能手机的诞生引领了 IT 技术的变革，全民从互联网时代跨入移动互联网时代。过去三年，中国网民得到了爆发式增长，截至 2015 年 6 月底，中国网民数达 6.68 亿，互联网普及率 48.8%；手机网民 5.94 亿，移动占比达 88%。

互联网科技与教育的结合，一方面使学习者获取资源变得得容易，同时也创造出了简单高效的学习方式。教育信息化作为国家高度重视的发展规划，互联网教育一方面受到政府的大力支持，另一方面也受到资本市场的热捧。

英语作为被世界上最多人使用的语言，不论是国际贸易中的商务交流还是地球村落里的社交沟通，一口流畅的英语口语变得越来越重要。面对目前国内无论是在校学生还是职场从业者普遍英语口语不佳的现状，很多教育科技企业推出了英语口语训练 App，本测评就是针对此类产品撰写的测评报告。

二、测评目的

为了检验市面上的英语口语测评软件是否真正具备口语测评实力，在几项语音测评技术上的表现力如何？互联网教育教育研究院几位软件测评师挑选了市面上具备口语测评能力的三款英语口语训练软件，在五项语音评测指标上进行了系统性测试并作出对照，希望可以为广大的英语口语学习者提供指导性意见。

三、测评指标

结合语音测评软件在实际应用过程中的权重比例，我们给不同的测评指标设置了不同的分数（总分为 100 分），如下图所示：

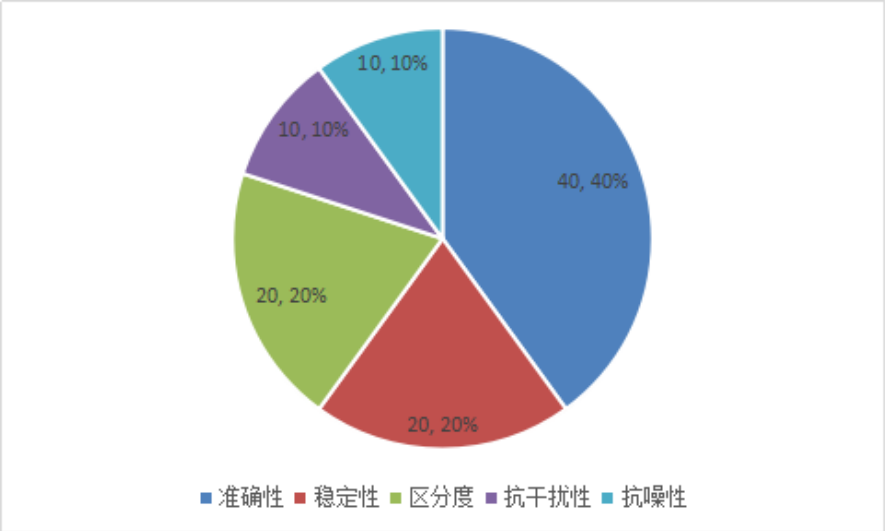

图 1 评分权重

指标	分值	说明
准确性	40 分	1.使用标准音，是否得高分。 2.准确度与得分是否呈线性相关。 3.能否区分句子中的单词读得好、中、差等。
稳定性	20 分	同一测试音，在不同测试，得分波动小。
区分度	20 分	是否区分口语内容中的无关成分（如汉语）
抗干扰性	10 分	1.排除快慢等无关干扰因素 2. 排除性别因素
抗噪性	10 分	能一定程度上排除周围环境噪音干扰得到较准确的测评结果。

表 1 产品评分表

1、准确性

准确性指的是口语测评软件在测试评价上拥有合理的打分标准。检验测评软件准确性最好的方式就是使用系统自带的原音去测试，如果能对原音打出高分，则说明测评软件具备很好的准确性，反之则准确性不高。

2、稳定性

稳定性指的是测评软件对同一测试内容在多次过程中所得的结果是否一致，或波动少波动幅度小。如果测试结果波动多幅度大，则该软件稳定性差。

3、区分度

区分度指的是测评软件对测试内容中无关成分的辨识能力。如果测评软件对朗读内容中的无关内容无法辨识并对其作出评价打分，则该软件区分度差。

4、抗干扰性

抗干扰性指的是测评软件能否排除测试内容中的无关干扰（如年龄因素，生理因素，发音快慢）因素得到准确测评结果的能力。如果面对无关干扰因素得到的测评结果偏离准确值，则说明该软件抗干扰性能差。

5、抗噪性

抗噪性是指测评软件实际应用过程中面对周围噪音环境获得较高准确结果的能力。如果在噪音环境中任能获得准确的测试结果，则抗噪性能好。反之，抗噪性能差。

四、测试环境

测评主要工作选择在约 15 平方米的独立会议室完成，以保证测试工作正常进行。在测试产品抗噪性能时，测试工作转移到户外环境下进行。测试中使用了 5 部手机，包括苹果、小米、魅族等常见手机品牌。

五、语音功能测评

1、准确性

准确性无疑是衡量口语测评技术最核心的指标之一，如果评得不准，无疑失去了作为测评标准的价值。

采用什么基准测试方法，才能有效测试出口语测评的准确性呢？经过一番思考，测评团队采用了原文标准发音作为测评是否准确的重要判断依据。在使用原文标准发音作为测试发音源时，如果测评技术奏效，应该能够得到很高的分数。如果原始的发音被判定为低分，则说明语音测评没能有效地区分高质量的发音，则可以认定不合格。下图为多次采用原音测试的不同产品的测试结果对照表。

内容	口语 100	同类产品 X	同类产品 Y
第 1 句	100	78	83
第 2 句	100	84	81
第 3 句	98	82	76
第 4 句	99	79	82
第 5 句	100	81	85

表 2 准确性测试对比

从对照结果可以看出，当使用原文相同的音源时，口语 100 的分值都很高，平均分在 99 分。但是在对其他口语测评软件进行标准音测试时，结果发现不同得分有很大的波动，说明口语 100 在评分算法上的确有过人之处。

2、稳定性

稳定性是衡量语音测评技术的第二大核心指标，一个在多次测评表现不稳定的语音测评产品是不具备信赖力的。

为了检验三款口语测评软件的稳定性，我们使用其系统自带的原音对话在两个手机上进行了多次测试。以下是其中 7 次测试结果的对照表。

内容	口语 100	同类产品 X	同类产品 Y
第 1 次	100	82	76
第 2 次	100	79	84
第 3 次	100	83	85
第 4 次	100	76	79
第 5 次	100	85	82
第 6 次	100	81	87
第 7 次	100	78	81

表 3 稳定性测试结果

从表格中我们可以看出，在多次测量结果中，口语 100 的得分保持了一致的高分。而另外两款产品均出现了明显波动。说明了口语 100 在语音测评上的稳定性要优于同类产品。

3、区分度

区分度是检验一款语音测评产品评测能力的基础指标，如果无法区分测试内容中的无关成分反为之打出了分数，则该产品就缺乏语音测评技术的专业水平。

为了检验三款英语口语练习软件在语音测评过程中能否区分无关的声音成分，我们使用AU 软件在完整的标准音中插入了鸟叫声、风吹声、水流声、木板敲击声、汉语声。用插入不同无关音的原音朗读分别对三款产品进行多次测试，测试结果如下。

无关声音成分	口语 100	同类产品 X	同类产品 Y
鸟叫声	无法打分	打低分	无法打分
风吹声	无法打分	打低分	打低分
水流声	无法打分	打低分	打低分
木板敲击声	无法打分	无法打分	无法打分
汉语声	无法打分	无法打分	无法打分

表 4 区分度测试结果

从检测对比结果中可以看出，在标准原声中插入无关声音，口语 100 完全能区分出来，并拒绝对该次测评打分且发出提示，其他两款产品对部分插入无关音的内容仍能打分。由此说明口语 100 在语音测评过程中表现出的专业性更强。

4、抗干扰性

抗干扰性是口语测评软件的基本技术指标，如果不能应对一些性别、年龄、快慢等无关因素得到准确的测评结果，该软件也就失去了应用的价值。

为了检验口语 100 的抗干扰能力，我们对标准原声使用调音软件做了性别变声和快慢的处理，分别对四款英语口语练习软件进行了抗干扰性测试。以下的测评结果对照表。

	口语 100	同类产品 X	同类产品 Y
原声	100	87	84
原声变女声	100	81	79
快 30%	99	76	72
蛮 30%	100	82	81

表 5 抗干扰测试结果

由测评结果对照表中可以得出，口语 100 在男女不同、快慢不同等无关干扰条件下表现的适应性很强，而另外两款产品对于这些无关干扰都表现出了差异性，即抗干扰能力弱。

5、抗噪性

抗噪性也是检验一款语音测评是否合格的关键指标，如果一款口语测评抗干扰能力差，将没办法在实际学习过程中使用。

为了检验口语 100 的抗噪能力，我们的测评师将测试手机带到了街边、公园和中学操场三个场景下。在不同的场景下使用标准音分别对三款软件测试了三次。以下是测试对照结果。

场景		口语 100	同类产品 X	同类产品 Y
街边	第一次	96	78	69
	第二次	95	74	72
	第三次	95	75	66

公园	第一次	99	81	79
	第二次	99	86	81
	第三次	98	83	83
学校操场	第一次	98	81	72
	第二次	98	79	77
	第三次	97	84	74

表 6 抗噪性测试结果

通过对三款软件在三种户外场景的多次测试中，我们发现口语 100 在这些噪音环境下任能保证较高的测评准确度。而另外两款软件在较吵的街边测试水准大幅降低，在少量噪音的公园中测试结果也出现了不少偏差。通过对出可以看出，在抗噪性能上口语 100 要优于另外两款软件。

六、其他功能测评

除了对三款口语练习 APP 的核心功能进行系统性测试，我们的测评师通过对软件的多次深度体验对其在技术和内容的多个层面也进行了测评。以下是点测评果对照表。

指标			口语 100	软件 X	软件 Y
技术（60 分）	界面设计(5 分)	界面设计是否美观（2 分）	1.4	1.4	2
		功能区域设计是否符合用户使用习惯（2 分）	2.6	2.6	2
		对不同设备是否有适配（1 分）	1	1	1
	操作性能(15 分)	启动速度（3 分）	2	2	2
		基本操作是否流畅（3 分）	2.2	2.6	2.4
		登录注册是否便捷（3 分）	2	2	2.2
		常用功能是否满足基本需求（3 分）	2.2	2.4	2.2
		课程播放下载是否流畅（3 分）	2.6	2.4	1.8
	语音功能（30 分）	语音识别功能准确度（30 分）	25.6	21	21.2
	系统稳定(10 分)	是否有闪退（5 分）	4.8	4.6	4.6
		操作无响应等严重 BUG（5 分）	5	4.2	4.5
内容（40 分）	专业性（10 分）	教学内容是否专业（5 分）	4.7	4.2	4

		口语发音是否标准（5 分）	4.4	4.2	4
	全面性（20 分）	教学内容全面丰富（10 分）	7	6.6	6.8
		能覆盖大多数用户的学习需求（10 分）	7.8	6.2	6.4
	即时性（10 分）	教学内容更新速度（5 分）	3.6	3.6	3.4
		教学内容结合时下热点（5 分）	4.2	3.8	3.8
合计			83.1	74.8	74.3

表 7 其他功能评分表

通过对测评数据的观察，我们发现在内容设置上口语 100 要优于另外两款软件。口语 100 在内容的专业性上更高，更能获得用户的信赖。在内容的全面性和即时性上口语 100 略好于另外两款软件，但从整体上看三者都有待提高，需要在内容上能更贴合用户需求。

在软件的技术性上，口语 100 与另外两款软件各有高低。体现在软件的界面设计和操作流畅度上，口语 100 的交互体验要略弱于另外两款软件，但在语音功能的准确度和系统稳定性上口语 100 表现更好。

七、测评总结

互联网教育研究院测评工作室通过对口语 100 和另外两款英语口语训练软件的产品测试，我们发现口语在 100 在核心测评功能上要优于其他软件，但在产品的交互体验和内容需求满足度上都有待提高。

核心功能指标上，口语 100 的准确性近乎完美，在与其他产品的对比上拥有绝对优势。在稳定性、区分度、抗干扰性和抗噪性四项指标上也要较优于另外两款产品，是一款专业性英语口语训练 APP。在对三款产品其他功能的测试中，口语 100 在界面设计和操作性能力要弱于另外两款产品，但在内容的专业性和全面性要要好于另外两款。相信口语 100 在满足用户需求的同时，能进一步改进用户体验，成为功能适用、学生爱用、家长认可的英语学习产品。