搜集可靠的评价数据

教师总是在进行着非正式的形成性评价。人类总有一种本能，不断地对周围的人和形势做出判断。然而，绝大多数这类判断都是无意识的，很多可能带来错误的印象和理解。为了使形成性评价中搜集的数据是有效的，就必须评估它需要评价的对象；为了使数据是可靠的，就必须提供该评价是可以复制的这样的信息。.

有效度的评价往往精确地指向具体的技能、策略和知识。如数学中要检测问题解决，但回答多项选择题并不能给予教师学生问题解决情况的信息。正确地回答这些问题只能说明学生记住了如何采取问题解决策略，或者说明他们非常善于猜测，但是却不能说明在真实的需要问题解决的情境中学生会怎样表现。这些很容易评分的评价对评价21世纪技能是没有效度的。

斯蒂金斯(Stiggins，2004)
警告说："我们并没有花费很多来保证课堂上评价的准确性，因此发生不准确的评价的几率以及随之而来的在所有层面上的错误决策明显地增加" (p. 25)
。当教师利用太少的信息对学生知识与能力太快做出决策时，他们的结论将抑制学生的成长而不是鼓励成长。

艾瑞森(Araison，2001) 描述了某些对效度产生的威胁：

1. 一成不变，根据自己的印象和原先的偏好来下结论。
2. 有逻辑上的错误，依据不相干的某些特征，如被强调的程度、其兄弟姐妹的成绩等来评价学生的成绩。这些判断常常是无意识的，教师们不觉得自己是这样做的。）
3. 依据不恰当的事例，只按照一次观察或片断的信息来做出判断。
4. 推而广之，推断学生在某一情境下的行为也会是其他情境下采取的行为。

搜集到的有关学生学业成绩的数据还必须是可靠的。可靠的信息往往是一致和典型的。任何关于学生思维评价数据，比方说在一个长假之前的某个日子搜集的，都有可能是不可靠的，因为这时学生的行为极有可能是非常态的。

由于评价数据是为了帮助教师得到有用的结论，因此它必须是有效度的，能说明某些东西是重要的，也必须是有信度的，说明某些东西是常态的。研究者会用"三角测量"这个术语来描述从数据得出结论的过程，就像在一个犯罪案件的证据被披露之前希望得到确证事实的新闻记者，
教师在对学生的能力得出结论之前需要不只一种信息。即便在这种情况下，结论也一定是暂定的，需要接纳有矛盾的数据。这意味着一位教师可能发现某个孩子在小组项目或写学习日志方面不会归纳，但是以后教师却发现她在其他学科领域是会归纳的。这样教师就可以暂且得出结论：这个孩子不会归纳是因为她还不具备足够的学科知识，而不是因为她缺乏思维的专业知识。

绝大多数教师都是很警觉的，持续关注他们的学生。他们虽然不能帮助，但是却会注意到学生们在做什么和说什么。不幸的是他们极少意识到这种非正式的观察是形成性评价，并且不会以一种系统的方式记录下他们所看到的东西。这些类型的观察，在没有仔细分析的情况下被使用的话，可能导致偏离真相的看问题视角或者错误的结论，因为它们未能考虑足够的数据。以偶然或者非系统方式搜集到的数据为基础的教学会阻碍学生的学习。从形成性评价中谨慎地搜集和考虑与学生相关的信息是很费时，并需要计划的，但是这类评价对学生学习和动机的影响让我们觉得这种努力是很值得的。