A Multi-Level/Multi-Type Model for Design-Based Alignment of Instruction, Assessment, and Testing

Daniel T. Hickey & Steven J. Zuiker
University of Georgia Learning & Performance Support Laboratory, 611 Aderhold Hall, Athens, GA 30602
Voice: 706-542-3157; Fax: 706-542-4321
dhickey@coe.uga.edu; szuiker@coe.uga.edu

Steven McGee
Center for Educational Technologies
Wheeling Jesuit University

Accountability-oriented reforms demand immediate *and* continual gains on criterion referenced achievement tests for all students, without diminishing other educational outcomes. The poster describes a model that builds on the *formative assessment* research of others, in order to meet this challenge. This model address three issues that expert panels have identified as needing resolution to maximize the potential of formative assessment: (1) different formats and contexts offer different formative potential; (2) the summative functions of assessment often undermine formative goals, and (3), potentially formative information is often unused or used in ways that undermines learning. The guiding principle of our approach is what theorists have called *systemic validity*, where the act of measuring student outcomes leads to increased performance on those same outcomes.

Four features work together to achieve systemic validity with this model. First, the model prioritizes assessment functions that *directly* advance learning. Thus, formative feedback provided directly to learners to advance their knowledge is given priority over feedback functions that advance learning indirectly (e.g., curricular refinement, remediation, grading, promotion, etc.). Second, this approach features *conversational feedback*. In contrast to a narrow "test-prep" focus on specific items, this discourse-based approach focuses on understanding the concept or skill behind a particular assessment item. Feedback is organized around learner-oriented *answer explanations* that detail the reasoning behind each item without directly stating the answer. Instead of corrosive considerations of prior performance, answer explanations scaffold discourse that advances current understanding of the assessed concept. Even rudimentary enactments of formative feedback (checking answers) guarantees some learning; more idealized (but attainable) enactments foster dramatic gains. The third feature in this approach is the use of multiple levels of assessment to balance formative and summative goals. Assessment theorists have identified five levels of summative assessment: *immediate, close, proximal, distal,* and *remote*. These roughly correspond with *informal observation, semi-formal classroom assessment, formal classroom assessment, criterion-referenced tests,* and *norm-referenced tests*. This model maximizes systemic validity by clarifying the unique formative and summative affordances of each level. Feedback conversations are introduced and ritualized at the more immediate levels, whose formative function can't be overlooked or undermined. This prepares students and teachers for powerful feedback conversations at the more distal levels that can most directly increase achievement test performance. Systemic validity is attained with multiple levels of aligned assessments: the formative value at one level can be refined using initial performance the next level (maximizing consequential validity); student performance at a third level then provides summative evidence of the impact of those refinements (maximizing evidential validity). With all five levels, this applies to all three levels of classroom assessment. The fourth feature is the use of "design-experimentation" to exploit the full potential of the multi-level model, maximizing the direct and indirect impact of formative feedback. Reflecting recent methodological advances, "intermediate-level" practical assessment theory is created within iterative cycles of refinement. Discourse analytic methods are used to refine assessment conversations and directly evaluate their impact on participation in domain knowledge practices; conventional experimental and quasi-experimental methods indirectly demonstrate impact to diverse audiences.

A three-year study reveals the potential of this model. Four hour-long close-level feedback conversations within roughly 20 hours of secondary genetics instruction yielded proximal gains of 1.98 SD and distal gains 1.06 (compared to .25 and .57 for matched comparison classes using conventional curriculum). New studies just getting underway include two and even three levels of feedback conversation and two levels of external tests.