

# Perspectives on UDL and Assessment:

## An Interview with Robert Mislevy

David T. Gordon, Jenna W. Gravel, Ed. M., &  
Laura A. Schifter, Ed. M.

National Center on UDL  
40 Harvard Mills Square, Suite 3 Wakefield, MA 01880-3233  
Tel: 781-245-2212  
Fax: 781-245-5212  
TTY: 781-245-9320  
Web: [www.udlcenter.org](http://www.udlcenter.org)

# **Perspectives on UDL and Assessment:**

An Interview with Robert Mislevy

David T. Gordon, Jenna W. Gravel, Ed. M., &  
Laura A. Schifter, Ed. M.

---

Paper / 3

---

April, 2010

## Editors' Note



Dr. Robert Mislevy is a leading expert in educational assessment, technology, and cognitive science. He was a Distinguished Research Scientist at Educational Testing Service (ETS), where he worked for nearly two decades. ETS develops, administers, and scores some 50 million standardized tests per year, including the SAT, AP, GRE, TOEFL and others. Dr. Mislevy is currently Professor of Measurement, Statistics, and Evaluation at the University of Maryland. He also collaborates with Cisco Learning Institute on advances in assessment. Dr. Mislevy holds a Ph.D. in Methodology of Behavioral Research from the University of Chicago. He was elected to the National Academy of Education in 2007.

Dr. Mislevy is well positioned to comment on UDL and assessment as co-Principal Investigator of [Principled Assessment Designs for Inquiry \(PADI\)](#). A joint project of the University of Maryland, CAST, and SRI International, with funding from the National Science Foundation, the PADI researchers aim to develop a framework for high quality assessments for science inquiry projects. In 2007, the team began working with CAST to infuse UDL principles into the assessment framework.

In this 2009 interview, Dr. Mislevy shares his insights on the historical background of assessment, the current state of assessment for students with disabilities, the integration of UDL and assessment, as well the implications for policymakers. He emphasizes the value of large-scale assessment but also notes that the field has not kept up with advances in the learning sciences. UDL is a way of applying those advances, and Dr. Mislevy points out that a principled application of UDL can in fact increase the value and even the validity of large scale assessment for a greater number of students.

## Interview

### How has the landscape of assessment changed over your career?

Over the past 30 years, advances in the field of psychology have given us a better understanding of how people learn and of the nature of knowledge. In the 1980s, we started to see the impact of information-processing and cognitive science in assessment. In the 1990s, socio-cognitive perspectives began to shape our understanding. Yet the funny thing is that the practice of large-scale testing really has not changed that much! Informal and local assessments have been changing over the years, but there have been few adjustments to the way in which we measure student learning on a large-scale. The 80-year-old perspective that a couple of questions answered over a couple of hours can give you an adequate understanding of what people know and what people can do persists in large-scale assessment. There are still reasons for doing some large-scale assessment. However, you no longer can assume that the little snippets from large-scale assessment tell you everything that is important to know. You get some information, but now it needs to be contextualized. And, you also need to understand what the information does not tell you as well as what it does.

What has changed with regard to large-scale assessment is the public's growing dissatisfaction with what we are learning from standard methodology. No Child Left Behind (NCLB) has spurred frustration with large-scale assessment because its accountability measures hang on these tests results. A good outcome of NCLB has been to direct more attention to the need for assessment and for knowing what kids are learning. But as [Harvard researcher] Dan Koretz's research suggests, the steps that you take to maximally improve NCLB scores are not necessarily the best steps that you would take to improve student learning, and reliance on test-score information alone can lead to bad policy.

### **What is the value of large-scale assessment for accountability?**

One of the values is what you might call “chain of custody.” Chain of custody is a term from the law that describes the chronological documentation of a piece of evidence. For example, after a piece of evidence is discovered at a crime scene, the names of all of the different people who examine the evidence and all of the protocols that are followed are recorded. This documentation enables you to make inferences from that evidence a year or maybe three years later. The evidence is “protected” so it can be used at times and places quite distant from its initial discovery.

Think, for example, about the SAT. The SAT still has value not simply because it is measuring people’s innate capabilities as a be all and end all, but because it serves as a little chunk of evidence. For a finite, reasonable amount of money, you know something about how evidence was collected, and you roughly know something about the chain of custody of that evidence. It is evidence that is gathered at locations throughout the country, summarized, and used by people at distant times and places. Despite its limits as evidence about what the students know and can do, it brings some credibility in terms of the chain of custody.

You sometimes hear the argument that we should not have large-scale, common assessments because teachers are better able to assess their students. It is claimed that teachers know their students better and that we should trust our teachers more in terms of their ability to conduct relevant, meaningful assessments that guide their instruction from day to day. Part of that argument holds water. In the local context, you can get more targeted and more useful information from students. However, such assessments—for different students, in different classrooms, and in different courses of study—while useful in local contexts, do not communicate value to people in distant times and places (policymakers, for example) who are removed several steps from the instructional episode. Ideally, you would really like to see large-scale assessments as working with rather than working against the local assessments that are used to determine how students learn, to obtain feedback, and to guide teachers’ instruction for the next day.

### **So perhaps large-scale assessment does not offer much in terms of improving classroom instruction and learning except as it can help lead to systemic changes or investments that in turn be used for local improvement?**

I would agree. However, this does not mean large-scale assessment could not be more informative for instruction. As an example, I would mention a project funded by the National Science Foundation that we are working on with our colleagues at CAST, SRI, and Pearson that focuses on the Minnesota Comprehensive Assessment Program. We’re working with large-scale science assessment at the middle school level. Minnesota’s assessment is computer administered, which allows them some flexibility in making the assessment more interactive. However, despite being computer administered, it is still traditional in that every kid in 8th grade in Minnesota looks at and interacts with the same test items. Our project is building design patterns around the science standards and benchmarks in Minnesota that can be used for both assessment and instruction. These design patterns help you build learning tasks, group work, and extended investigations in the classroom that are compatible with the same state standards and benchmarks. And, because the same design patterns are also used in assessment, the content included on the large-scale test, the Minnesota Comprehensive Assessment II (MCA-II), is made up of little snippets from the design patterns that teachers are using in their classrooms. By using the same design patterns for both instruction and assessment, teachers and curriculum developers begin to understand what the standards and benchmarks are really about and the various ways to teach them and assess them in the classroom. When the MCA-II test comes at the end of the year, teachers can be assured that there will most likely be some tasks included in the assessment that are like are similar to the activities that have been taking place in the classroom all year long. Students are going to do better on the items, not because they studied and focused just on the MCA-II kinds of tasks, but because they learned the ways of thinking that are compatible with what you do on those MCA-II tasks in the simpler forms that can be presented there. This is just one example of how the large-scale testing can be compatible with learning goals without subverting teaching.

**In an ideal world, how would large-scale assessment and other measures, such as curriculum-based measurement or progress monitoring, integrate to give us reliable measures for accountability purposes?**

To answer this question, I will give you another example from industry: the Cisco Networking Academy, a global e-learning initiative that focuses on developing introductory skills in network engineering. Cisco has designed simulation-based tasks for courses that deal with troubleshooting, network design, and setting up networks. The simulation tasks range in their level of complexity. Some may require students to work collaboratively for a number of hours, while less complex tasks may require only ten minutes of individual work. The final exams are based on smaller segments of the same kinds of simulation tasks. At the end of the course, students are well prepared for the exam because the assessment is compatible with the ways in which they have been learning—and the skills that they need in the real world. In fact, their focus of study and their final exam will also be compatible with some of the tasks on their certification test should students decide to pursue the area further. The final exams are not developed simply from a measurement point of view; they are derived from the same learning goals and the same psychology of learning used to create the courses.

The Cisco case also shows how technology can help to make congruency between instruction and assessment even stronger. Technology is loosening the constraints we have experienced over the eighty years of large-scale testing, and these constraints are being loosened in ways that make it easier to connect what happens in learning to what happens in assessment. For example, if you were conducting large-scale testing on the computer, students could build interactive models to solve problems, they could design their own graphics, or they could map information from one form of representation to another. Technology removes the layer of artificiality that develops from the tight constraints of traditional large-scale assessments.

**What is the current state of assessment for students with disabilities, English language learners, and others marginalized in traditional curriculum? Do you believe we are currently able to accurately measure the learning of all students?**

I am not an expert in everything that is being done in the field regarding assessing students with disabilities, but I have done a lot of thinking around the challenges of traditional assessment. The traditional way of building assessment arguments has always been to present everyone with the same stimulus material and have them interact and respond in the same ways. It was believed that this traditional method allowed you to obtain comparable evidence that could be interpreted in the same way for everyone who is tested. However, there is a big problem with this approach; it forces you to make assumptions about students' ability to see, to hear, to physically interact with their surroundings, to concentrate, and so on. Although many people realized that including students with these types of varying abilities would lead to invalid inferences, the students were often tested nonetheless because the alternative was to exclude them from the assessment all together. The real challenge is finding ways to make assessment arguments when different students might be interacting with different stimuli and responding in different ways. The traditional educational measurement and assessment machinery, conceptual as well as statistical, does not help us to address this issue. However, I hope that the work that is being done in the field will lead us to some answers.

The PADI project seeks to lay out a rigorous conceptual basis for assessment using principles for Universal Design for Learning (UDL). We are working to develop more tailored assessment situations for different individuals that still make sense using the same frame of interpretation across individuals. That frame of interpretation is driven by clearly defining the specific knowledge, skills, or abilities that we want our students to develop. Often times, a "one size fits all" assessment is developed, and adaptations and accommodations are made only after the fact. This is the hard way to build accessible assessments; the adaptations are constrained by the "one size fits all" conception of how you develop and interpret tests. This method does not tell you how to think in a principled way about the students for whom the "one size fits all" test does not work. After-the-fact adaptations might make some improvements in inclusion, but you do not put it in the framework for drawing the inferences that you want to make. In the PADI project, we are thinking about the range of students' abilities and disabilities from the very beginning. It is a different way of thinking about what assessment is.

**What kind of a role do you see UDL playing in assessment?**

UDL prompts you to target learning goals; you identify what we call the “focal knowledge, skills, and abilities” or “focal KSAs,” that you want your students to develop. When applying UDL to assessment, you are evaluating these focal KSAs in order to determine if students are making progress in those capabilities.

UDL also encourages us to carefully consider all of the knowledge, skills, or abilities that might tangentially be involved in assessing the focal ones. These “non-focal KSAs” might prevent students from accurately being able to demonstrate what they know and what they can do. For example, students with a visual impairment might do poorly on a science assessment not because they do not know the content but because they are unable to see the material. Other students may do poorly on a specific item simply because they were not given some construct-irrelevant information that they would need to know in order to interact with the task. In both of these examples, non-focal KSAs interfere with students’ learning and performance on tests, and lead to invalid assessment. UDL pushes us to think about the ways in which we can support students’ non-focal KSAs so that we can target and address the actual learning goals.

**Questions have arisen about the impact of UDL upon the validity of assessment. From your perspective, would applying UDL to large-scale assessment have a negative impact upon construct validity?**

The historical conception of construct validity was built for the situation in which “everybody gets the same test and does the same thing so that we have the same data.” The conceptual and statistical machinery for thinking about validity was built upon this framework as well. However, if you apply UDL to assessment and assess different kids in different ways, the procedures and the methods that went along with ensuring validity based upon that framework no longer apply. Integrating UDL into assessment takes away all of the tools for thinking about validity that we have relied upon in the past. It requires a different way of thinking.

The application of UDL to assessment has the potential to either increase or decrease construct validity. A straight ad hoc application of UDL without identifying the focal knowledge and skills that are being assessed will most likely lead to a decrease in construct validity. However, if UDL is applied in a principled manner, it will actually increase construct validity for a larger population of students. Our PADI project focuses upon ways of growing the population of students for whom we can obtain valid measures. We are working to develop tools and procedures that make the application of UDL explicit and that help us make a rigorous assessment argument for when different students might be presented tasks in different ways, or interact with them or respond to them in different ways, to be able to get at the focal knowledge and skill. We want to help the field to understand that a principled application of UDL can increase construct-validity for a bigger population of kids. But simply applying the UDL principles in and of itself is no guarantee.

**What are some ways in which policy can support this more thoughtful, more “principled” application of UDL to large-scale assessment?**

There is a big challenge in implementing policy around assessment. Many people believe that they know everything there is to know about teaching and about testing because they have been taught and because they have taken tests. That sort of “every day thinking” about testing will not bring to bear what we are learning from cognitive science nor will it support the integration of UDL and assessment. Policies that are made using eighty year -old ways of thinking about learning and assessment can sometimes make the job of improving assessment more difficult.

Policy can play a role in instituting the UDL framework as an ingrained part of the assessment design process. Creating tasks that have this principled reasoning throughout them needs to be a part of development from the very beginning. At first, this new approach to design will require more thought and more money for test development, administration, and review. For eighty years, tests have been thought of as “one size fits all,” and it has been acceptable to make that “one size” as fast and as cheap as possible. This way of thinking has left a bad legacy. The notion that what you can do fast, cheap, and large scale is in fact sufficient for everybody still persists, often at the policymaker level. Recognizing that test development requires thoughtful consideration and significant funding would

be a very big step forward. It is true that this new thinking around assessment design takes a little more money, and the procedures for creating and administering tests are a little more involved. However, there are ways to make this assessment design process more economically feasible. The use of technology and the sharing of design patterns and assessment delivery mechanisms across states and across testing companies will save on time and resources. In summary, policymakers can do two things to promote assessment for a wider range of students: 1) they can make it a priority to understand the current barriers that are preventing us from obtaining accurate measures of learning for a broad range of students, and 2) they can support the development of new technologies, design patterns, and assessment delivery mechanisms that can be shared by all.

Furthermore, it is important to promote examples of successful initiatives in order to generate support for this new way of thinking around assessment. Many other sectors are using more innovative assessment and progressing at a much faster rate than our public schools. In the medical profession, for example, the use of simulations has made a positive impact upon both assessment and learning. The National Board of Medical Examiners has now included computer case management problems as part of the licensure sequence. In medical schools, students continue to memorize the names of the bones, and muscles, etc., but the use of simulations encourages them to spend more time thinking through how that knowledge applies to interactions with patients, to diagnosis, and to treatment. The military provides another example. The military places a real value in having people learn to do things better. They have the motivation to improve instruction and assessment as best they can because the consequences are so real. The defense community is bureaucracy as much as any school district, but the military has the impetus to be creative in their approach to learning and assessment. Finally, community colleges seem to be a bit more flexible, innovative, faster to change. The lay of the land seems a little more favorable in community colleges than it does in K-12 or in university systems. Publicizing examples of success such as the Cisco Networking Academy, the use of simulations in medical schools, and the innovation within the military and many community colleges will help the public to understand that this new way of thinking about assessment is actually possible. Once parents, teachers, and citizens see the effectiveness of these examples, they will become a little less tolerant of the policymakers who are not working to move these developments forward.

THE NATIONAL CENTER ON UDL AT  **CAST**

40 Harvard Mills Square, Suite 3, Wakefield, MA 01880-3233  
(781) 245-2212 • [udlcenter@cast.org](mailto:udlcenter@cast.org)