1

Introduction to Handbook

André A. Rupp and Jacqueline P. Leighton

COPYRIGHTED MATERIAL

Motivation for Handbook

The field of educational assessment is changing in several important ways at the time of this writing. Most notably, there has been a shift to embracing more complex ways of thinking about the relationship between core competencies, behaviors, and performances of learners at various developmental levels across the lifespan. These new ways of thinking have been fueled by new models of cognition that are increasingly more inclusive and accepting of correlates of basic knowledge and skill sets. In many educational assessment contexts, considerations of how cognitive, meta-cognitive, socio-cognitive, and noncognitive characteristics of individual learners affect their individual behaviors and performances – and those of teams that they are working in – are becoming increasingly common. Clearly, at a basic level, the mere conceptual consideration of such broader characteristics and their interrelationships is not intellectually new but the way in which they are nowadays explicitly articulated, operationalized, and used to drive instructional and assessment efforts is indeed something new.

Assessment of Twenty-First-Century Skills

In US policy, this trend is reflected in curricular movements such as the *Common Core* and its adoption by individual states as well as collections of states in consortia such as the *Partnership for Assessment of Readiness for College and Careers* and *Smarter Balanced*. While the degree of influence of these two particular consortia is likely to change over time, the foundational tenets and goals of the *Common Core* are less likely to vanish from our educational landscape. Importantly, *Common Core* standards articulate models of learning that are explicitly focused on the longitudinal development

The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications, First Edition.
Edited by André A. Rupp and Jacqueline P. Leighton.
© 2017 John Wiley & Sons, Inc. Published 2017 by John Wiley & Sons, Inc.

of learners over time across grades. Focal competencies include domain-specific knowledge, skills, and abilities as well as professional practices but also broader cross-domain competencies.

Such complex competencies are sometimes called "twenty-first-century skills" and include cognitive skills such as problem-solving, systems thinking, and argumentation skills, intrapersonal skills such as self-regulation, adaptability, and persistence, as well as interpersonal skills such as collaboration skills, leadership skills, and conflict resolution skills. Of note is the inclusion of information and communication technology skill sets, which are an integral part of the digitized life experiences of citizens in our times across the world. As a result, the kinds of intellectual and creative tasks that effective citizens need to be able to solve nowadays with digital tools are often qualitatively different in important ways from the tasks of the past. As a result, considerations of smart assessment design, delivery, scoring, and reporting have become much more complex.

On the one hand, more "traditional" assessments constructed predominantly with various selected response formats such as multiple-choice, true-false, or drag-and-drop are certainly here to stay in some form as their particular advantages in terms of efficiency of scoring, administration, and design are hard to overcome for many assessment purposes. This also implies the continued administration of such assessments in paper-and-pencil format rather than digital formats. While it clearly is possible to use tools such as tablets, smartphones, or personal computers for the delivery of innovative digital assessments, many areas of the world where education is critical do not yet have access to reliable state-of-the-art technological infrastructures at a large scale.

On the other hand, there are numerous persistent efforts all over the world to create "smarter" digital learning and assessment environments such as innovative educational games, simulations, and other forms of immersive learning and assessment experiences. Sometimes these environments do not proclaim their assessment goals up front and may perform assessment quietly "behind-the-scenes" so as to not disturb the immersive experience – an effort called "stealth assessment" by some. Since the tasks that we create for learners are lenses that allow us to learn particular things about them and tell evidence-based stories about them, we are nowadays confronted with the reality that these stories have become more complex rather than less complex. This is certainly a very healthy development since it forces assessment design teams to bring the same kinds of twenty-first-century skills to bear to the problem of assessment systems development that they want to measure and engender in the learners who eventually take such assessments.

Methodologies for Innovative Assessment

In the most innovative and immersive digital environments the nature of the data that are being collected for assessment purposes has also become much more complex. We now live in a world in which process and product data – the indicators from log files that capture response processes and the scores from work products that are submitted at certain points during activities – are often integrated or aligned to create more comprehensive narratives about learners. This has meant that specialists from the discipline

of psychometrics have to learn how to play together – in a common and integrated methodological sandbox – with specialists from disciplines such as computer science, data mining, and learning science.

Integrating disciplinary traditions. Clearly, professionals deeply trained in psychometrics have a lot to offer when it comes to measuring uncertainty or articulating evidentiary threads for validity arguments when data traces such as log files are well structured. Similarly, professionals deeply trained in more predominantly computational disciplines such as computer science or educational data mining have a lot to offer when it comes to thinking creatively through complex and less well-structured data traces. Put somewhat simplistically, while traditional psychometrics is often seen as more of a top-down architecture and confirmation enterprise, modern computational analytics is often seen as a more bottom-up architecture or exploration enterprise.

In the end, however, most assessment contexts require compromises for different kinds of design decisions and associated evidentiary argument components so that effective collaboration and cross-disciplinary fertilization is key to success for the future. This requires a lot of strategic collaboration and communication efforts since professionals trained in different fields often speak different methodological languages or, at least, different methodological dialects within the same language.

Paradoxically, we are now at a time when conceptual frameworks like assessment engineering or evidence-centered design – a framework that many authors in this *Handbook* make explicit reference to – will unfold their transformational power best, even though some of them have been around in the literature for over 20 years. None of these frameworks is a clear "how-to" recipe, however. Instead, they are conceptual tools that can be used to engender common ways of thinking about critical design decisions along with a common vocabulary that can support effective decision-making and a common perspective on how different types of evidence can be identified, accumulated, and aligned.

Integrating statistical modeling approaches. Not surprisingly perhaps, the statistical models that we nowadays have at our disposal have also changed in important ways. Arguably there has been a strong shift in the last decades toward unification of statistical models into coherent specification, estimation, and interpretation frameworks. Examples of such efforts are the work on generalized linear and nonlinear mixed models, explanatory item response theory models, and diagnostic measurement models, to name just a few. Under each of these frameworks, one can find long histories of publications that discuss individual models in terms of their relative novelties, advantages, and disadvantages. The unified frameworks that have emerged have collected all of these models under common umbrellas and thus have laid bare the deep-structure similarities across these seemingly loosely connected models.

This has significantly restructured thinking around these models and has helped tremendously to scale back unwarranted, and rather naïve, claims from earlier times about the educational impact that certain kinds of statistical models could have by themselves. Put differently, it has helped many quantitative methodologists to re-appreciate the fact that any model, no matter how elegantly it is specified or

estimated, is, in the end, just a technological tool. Like any tool, it can be used very thoughtfully as a "healthy connective tissue" for evidence or rather inappropriately leading to serious evidentiary "injuries."

Integrating assessment design and validity argumentation. From a validity perspective, which is foundational for all educational assessment arguments, the constellation of design choices within an assessment life cycle has to be based on sound scientific reasoning and has to rhetorically cohere to provide added value to key stakeholders. This typically means that the information that is provided from such assessments should provide real insight into learning, performance, and various factors that affect these.

As such, smart assessment design considers the system into which the assessment is embedded just as much as the tool itself. In fact, under views of the importance of measuring learning over time as articulated in the *Common Core*, for instance, it is impossible to think of the diagnostic process as a one-off event. Instead, assessment information needs to be interpreted, actions need to be taken, experiences need to be shaped, and new information needs to be collected in an ever-continuing cycle of learning, assessment, and development. In this new world of cognition and assessment such a longitudinal view will become more and more prevalent thus forcing many communities of practice to change the way they design, deliver, score, report, and use assessments.

This perspective critically affects the societal reverberations that assessments can have when serving underrepresented or disadvantaged groups in order to improve the life experiences of all learners across the societal spectrum and lifespan. It may certainly be too much to ask of measurement specialists – or at least it may be rather impractical for workflow considerations – to always keep the bigger philanthropic goals of assessments in mind as these do not always influence their work directly. For example, the optimal estimation of a complex latent variable model for nested data structures will not be directly affected by an understanding of whether this model is used in an assessment context where assessment scores are used to provide increased access to higher-education institutions for minorities or in an educational survey context where they are used for accountability judgments.

However, ensuring that assessment arguments are thoughtful, differentiated, and responsible in light of societal missions of assessment is important, especially in interdisciplinary teams that are charged with various critical design decisions throughout the assessment lifecycle. It will help these teams be more motivated to keep track of controversial design decisions, limitations of assessment inferences, and critical assumptions. In short, it will help them to make sure they know what evidence they already have and what evidence still needs to be collected in order to support responsible interpretation and decision making. As mentioned earlier, such a shared understanding, perspective, and communal responsibility can be fostered by frameworks such as assessment engineering or evidence-centered design.

Integrating professional development and practical workflows. These last points speak to an aspect of assessment work that is often overlooked – or at least not taken as seriously as it could – which is the professional development of specialists who have to work in interdisciplinary teams. There is still a notable gap in the way universities train graduate students with Master's or PhD degrees in the practices of assessment design, deployment, and use. Similarly, many assessment companies or start-ups are under immense

business pressures to produce many "smart" solutions with interdisciplinary teams under tight deadlines that take away critical reflection times.

In the world of *Common Core*, for example, short turnaround times for contracts from individual states or other stakeholders in which clients are sometimes asked to propose very complex design solutions in very short times can be problematic for these reflection processes. While short turnaround times would be feasible if the needed products and solutions truly fit a plug-and-play approach, the truth is that the new assessment foci on more complex, authentic, collaborative, and digitally delivered assessment tasks require rather creative mindsets. They also require new modes of working that go from a simple design-and-deploy approach, interspersed with one or two pilot studies and a field trial, to a much more consistent design-deploy-evaluate-revise lifecycle with shorter and more frequent bursts of activity, at least for formative assessments. These mindsets require time to cultivate and established processes require time to change, which is again why frameworks like assessment engineering and evidence-centered design can be so powerful for engendering best practices.

Handbook Structure

In the context of all of these developments it became clear to us that it would not be possible to create a single *Handbook* that would be able to cover all nuances of assessment and cognition, as conceived broadly, in a comprehensive manner. Instead, what we have strived to do is to provide a reasonably illustrative crosswalk of the overall landscape sketched in this brief introduction. We did so with an eye toward taking stock of some of the best practices of the current times while setting the stage for future-oriented ways of rethinking those best practices to remain cutting-edge. After some back-and-forth we eventually decided to divide this *Handbook* into three core parts even though readers will find a lot of cross-part references as many ideas are clearly interrelated. For simplicity of communication, we decided to label these three parts *Frameworks*, *Methodologies*, and *Applications*.

Frameworks

In the *Frameworks* section we invited authors to articulate broader ways of thinking around what models of cognition might offer in terms of the psychological infrastructure that sustain frameworks for assessment design, delivery, scoring, and decision making along with associated validation practices. This part, in many ways, is a conceptual cornerstone for any and all types of assessments that are primarily developed with the intention to support claims about the unobservable information processes, knowledge, and skills that accompany observed performance. The nine chapters in this part present distinct but overlapping perspectives on how models of cognition can inform – both conceptually and practically – the design and developments of assessments from start to finish.

In Chapter 2 on the role of theories of learning and cognition for assessment design and development, Nichols, Kobrin, Lai, and Koepfler present a framework and three criteria for evaluating how well theories of learning and cognition inform design and

decisions in principled assessment design, assessment engineering, and evidence-centered design. In Chapter 3 on cognition in score interpretation and use, Ferrara, Lai, Reilly, and Nichols further analyze the elements that define principled approaches to assessment design, development, and implementation before comparing and illustrating the use of different approaches. In Chapter 4 on methods and tools for developing and validating cognitive models in assessment, Keehner, Gorin, Feng, and Katz focus us on ways to characterize cognitive models, including the rationale for their development and the evidence required for validation so as to ensure their utility for meeting assessment goals. This includes clearly defined assessment targets, a statement of intended score interpretations and uses, models of cognition, aligned measurement models and reporting scales, and manipulation of assessment activities to align with assessment targets, all within a backdrop of ongoing accumulation and synthesis of evidence to support claims and validity arguments.

In Chapter 5 on an integrative framework for construct validity, Embretson illustrates how a cognitive psychological foundation for item design and development can not only influence reliability but also the five aspects of an integrated construct validity framework with special attention on how automatic item generators are supported within the context of the framework. Further expanding on this idea, in Chapter 6 on cognitive models in automatic item generation, Gierl and Lai similarly show us how cognitive item models can be operationalized to guide automatic item design and development to measure specific skills in the domains of science and medicine.

In Chapter 7 on social models of learning and assessment, Penuel and Shepard analyze ways in which research teams articulate the vertices of the "assessment triangle." This includes representations of how students become proficient in the domain, the kinds of activities used to prompt students to do or say things to demonstrate proficiency, and frameworks for making sense of students' contributions in these activities in ways that can inform teaching. In Chapter 8 on socio-emotional and self-management variables in assessment, Kyllonen explains the importance of noncognitive skills as predictors of cognitive skills development and as outcomes for which assessments should be developed for their own sake. In chapter 9 on the role of cognitively-grounded assessment practices in understanding and improving accessibility for special populations, Ketterlin-Geller outlines the ways in which educational assessments can be enhanced in their design and development to be accessible to students in special populations. Finally, in Chapter 10 on integrated perspectives of validation and automated scoring, Bejar, Mislevy, and Zhang discuss the various design decisions that have to made during the lifecycle of automated systems for scoring and feedback. They specifically discuss the history of certain key systems across a wide variety of domains with applications that span short and extended written responses, spoken responses, responses with multimodal outputs, and interactive response processes within virtual learning environments.

Methodologies

In the *Methodologies* section we asked authors to present statistical modeling approaches that illustrate how information about cognitive processes can be operationalized and utilized within the context of statistical models. One potential conceptual

dimension to draw between modeling approaches is that of parametric versus nonparametric modeling approaches. The former are generally characterized by explicit functional forms, which include parameters that can be interpreted, strong assumptions that are made about distributions of component variables for estimation, and a variety of computational approaches for obtaining parameter estimates given suitable data. These models allow for the power of formal statistical inference around these parameters so that interpretations about cognitive processes or behaviors in the population can be made with the sample data. This particular quantification of statistical uncertainty is unique to parametric models even though there are other ways of quantifying uncertainty in nonparametric approaches. Moreover, parametric models allow for an explicit assessment of model-data fit using the parameters in the model and can be used efficiently for applications that require modularity and component-based information such as computer-adaptive (diagnostic) assessment, automated item generation, automated form assembly, and the like.

Nonparametric approaches are generally characterized by weaker distributional assumptions and use either probabilistic or rule-based decision sequences to create data summaries. While the focus of inference may be similar as with parametric models, the kind of information obtained from these models and the way that one can reason with that information is thus structurally distinct. For example, diagnostic measurement models and clustering approaches can both be used to sort learners into unobserved groups. However, in the former parametric approach one obtains parameters that can be used explicitly to characterize the learners and the tasks that they were given. In the latter nonparametric approach, such characterizations have to be made through various secondary analyses without explicit model parameters as guideposts.

The formalism of parametric models is certainly important whenever assessments are administered at larger scales and when decisions take on a more summative nature, perhaps for state-wide, regional, national, or international accountability purposes. However, the power of parametric models can sometimes also be useful in more formative decision-making contexts such as digital learning and assessment environments that require certain kinds of automation of evidence identification and accumulation procedures. Consequently, the six chapters in Part II of the *Handbook* are skewed more toward the parametric space overall, which is arguably appropriate given how powerful and important this model space is for educational assessment.

In Chapter 11 on explanatory item response theory models, De Boeck, Cho, and Wilson discuss how to specify, estimate, and reason within a unified latent-variable modeling framework called explanatory item response theory. The general idea is that this framework subsumes simpler modeling approaches from item response theory, which are the current state-of-the-art for data modeling in large-scale assessment. However, they expand upon these foundations by allowing for the inclusion of additional variables – called covariates – for learners, tasks, or learner-task combinations that may help to "explain" observed performance differences. As with any statistical methodology, the degree to which such explanations are robust and defensible more broadly based on scientific grounds requires additional validation studies. In Chapter 12 on longitudinal latent-variable models for repeated measures data, Harring and Houser discuss how to specify, estimate, and reason within another unified latent-variable modeling framework that focuses on the modeling of data collected over time

or other conditions of replication. They describe how seemingly complicated design choices in mathematical structures of certain model components can be – and have to be – grounded in an understanding about cognitive processes in order to make interpretations defensible. As with explanatory item response theory models, this framework allows for the inclusion of various covariates at the learner, task, or occasion level with similar evidentiary requirements for thorough validation of interpretations.

In Chapter 13 on diagnostic classification models, Bradshaw discusses how to specify, estimate, and reason with yet another unified latent-variable modeling framework called the log-linear cognitive diagnosis model. The general idea here is that an a priori specification of how different tasks measure different skill sets can be used to create classifications of learners into different competency states that are describable through these skill sets. Just as in the other two chapters discussed previously, covariates at different levels can be included into these models for additional explanatory power. In Chapter 14 on Bayesian networks, González-Brenes, Behrens, Mislevy, Levy, and DiCerbo describe how to specify, estimate, and reason with a family of latent-variable models that share many similarities, but also display critical differences, with diagnostic classification models. Similar to the latter models, these models require an a priori specification of relationships between skill sets and tasks, which can be refined through model-data fit evaluations. However, in contrast to those models, all the variables in this approach are categorical, the specification of relationships between variables can accommodate a large number of dependencies relatively easily, and the estimation is very general and well aligned with conceptual understandings of how human beings reason more generally.

In Chapter 15 on the rule-space methodology and the attribute hierarchy method, Cui, Gierl, and Guo describe a predominantly nonparametric alternative to diagnostic classification models and Bayesian networks. Specifically, their two methods represent historical foundations for the parametric approaches and remain attractive alternatives in situations where the full power of parametric inference is not needed. Both methods are used predominantly for classifying learners, with less of an emphasis on obtaining detailed characterizations of tasks or explanatory narratives through additional covariates, at least not within a single estimation run. Finally, in Chapter 16 on educational data mining and learning analytics, Baker, Martin, and Rossi provide an overview of the utility of a variety of statistical analysis techniques in the service of performing cognitively grounded data mining work for assessment purposes. They illustrate this work through applications in innovative digital learning environments where a wide variety of behavior detectors have been used to characterize learner actions and to make inferences about underlying cognitive skill sets and meta-cognitive factors that affect performance. This last chapter serves as somewhat of a conceptual bridge between the *Methodologies* and the *Applications* parts of the *Handbook* as the latter part contains more such innovative applications along with slightly more traditional ones.

The six chapters in this section clearly do not cover the entire space of psychometric or computational techniques that could conceivably be brought to bear to model observable learner behavior and task performance in order to make inferences about certain cognitive correlates. Entire books have been written about each of the modeling approaches, both within disciplines and across disciplines, which make any claim to a truly comprehensive coverage prohibitive. For example, we could have included

chapters on structural equation models or traditional item response theory models as well as chapters on other nonparametric clustering techniques or multivariate analysis methods.

However, it was not our goal to develop yet another methodological *Handbook* that is oriented primarily toward specialists whose day-to-day job is to make smart decisions about data analysis. Instead, we wanted to create a meaningful cross-section of this broad methodological space in a way that gives explicit room for arguments about how to specify, estimate, and, most importantly, reason with these models. We made strong efforts to work with the authors to keep the chapters in a rather accessible language, structure, and level of detail so that specialists who do not think about statistical models on a daily basis would be able to learn a few meaningful and actionable pieces of information about these methodologies from the chapters. It is our firm belief that even a tentative understanding and an associated thirst to learn more about the strengths and limitations of different modeling approaches can go a long way toward fostering this shared methodological and evidentiary reasoning understanding that we have talked about at the outset.

Applications

In the *Applications* section we asked authors to traverse an equally diverse space of possible uses of models for cognition in the service of a broad range of assessment applications. For example, we decided to select a few very common assessment applications and encouraged the authors of the seven chapters in this part to describe both the broader contexts and frameworks within which their illustrations are embedded and to be forward-thinking in their description. That is, rather than asking them to merely describe the state of the world as it is now we explicitly wanted them to take some intellectual chances and speculate on what some key trends for their areas of work would be.

In Chapter 17 on large-scale standards-based summative assessments, Huff, Warner, and Schweid discuss how thinking about cognition influences the design and use of these kinds of assessments. They use three powerful examples across different use contexts to show surface-level differences and deep-structure similarities across these contexts using a recent framework for differentiating between cognitive models. Using these examples, they articulate how certain kinds of articulations and operationalizations of cognition are necessary to increase the inferential power of these assessments and how others can be quite harmful to this process as they are somewhat unrealistic – or poorly matched – in this context. In Chapter 18 on large-scale educational surveys, Oranje, Keehner, Persky, Cayton-Hodges, and Feng discuss the general aims of these kinds of assessments, which is accountability at state or country levels, and illustrate the current innovation horizon in this area through examples from an interactive national assessment in the United States. They demonstrate that historical notions of item type restrictions are only partly transferrable for the future of this line of work, and that more complex interactive assessment tasks are the generative framework that should be utilized to measure at least some twenty-first-century skill sets reliably at this level of assessment.

In Chapter 19 on professional certification and licensure examinations, Luecht provides practical examples to show why assessment engineering design components

and procedures, including task modeling, task design templates, and strong statistical quality control mechanisms, are an integral and important part of the many processes for developing cognitively based formal test specifications, building item banks, and assembling test forms that optimize professional knowledge assessment and/or skill mastery decisions. In Chapter 20 on the in-task assessment framework for in-task behavior, Kerr, Andrews, and Mislevy describe an articulation of the evidence-centered design framework within digital learning and assessment environments specifically. They describe a set of graphical tools and associated evidentiary reasoning processes that allow designers of such environments to make explicit the different steps for operationalizing construct definitions for complex skill sets. These tools then help to link observable behaviors captured in log files to different construct components to derive useful feedback and scores that are based on an explicit chain of evidence, a process that they illustrate with three examples from different domains.

In Chapter 21, on digital assessment environments for scientific inquiry skills, Gobert and Sao Pedro provide yet another application of cognitively inspired assessment – in this case, it is the design, data-collection, and data-analysis efforts for a student-based digital learning and assessment environment devoted to scientific inquiry and practices. In Chapter 22, on stealth assessment in educational video games, Shute and Wang look at how both commercial games and games designed or adapted for assessment purposes can be powerful levers for measuring twenty-first-century skills. They describe how evidence-centered design thinking coupled with systematic synthesis of the current cognitive literature on these skill sets are necessary prerequisites for instantiating best evidentiary reasoning practices through embedded assessment in these contexts. In Chapter 23 on conversation-based assessment, Jackson and Zapata-Rivera introduce us to the benefits of these kinds of assessment for collecting new types of explanatory evidence that potentially afford greater insight into test taker cognition and metacognition. They further propose a new framework to properly situate and compare conversation-based assessments with other kinds of assessment items and illustrate the power of conversation-based assessment through a prototype. Finally, the Handbook contains a glossary with definitions of key terms that are used across chapters. In each chapter, the first mention of any key term in the glossary is boldfaced for easy reference.

Closing Words

As this brief overview has underscored, the *Handbook* that you are holding in front of you is a complex labor of love that involved the participation of many wonderful members of scientific communities engaged in some type of educational assessment activity. These activities span the design of large-scale educational surveys, the development of formative learning systems, the evaluation of novel statistical methods that support inferences, and the conceptual articulation of frameworks that guide best practices, to name a few. We are infinitely grateful for all of our colleagues who have worked patiently with us to create our particular conceptual crosswalk of this landscape. We sincerely hope that the final product will be as much appealing to them as it is to us.

Most importantly, however, we sincerely hope that readers will find this *Handbook* powerful for changing the ways they think about the interplay of assessment and cognition. We hope that reading individual chapters, parts, or maybe even the entire

book will stimulate new ideas, new ways of thinking, a thirst for wanting to learn more from references that are cited, and a deep continued passion for improving the lives of learners across the world through thoughtful and innovative assessment design, development, deployment, and use. If we were to make even small but meaningful contributions to these efforts we would be eternally grateful.

Sincerely,
André A. Rupp and Jacqueline P. Leighton