Models, Metrics, and Measurement in Developmental Psychology

Zachary Stein and Katie Heikkinen

Abstract: Developmental psychology is currently used to measure psychological phenomena and by some, to re-design communities. While we generally support these uses, we are concerned about quality control standards guiding the production of usable knowledge in the discipline. In order to address these issues precisely, we provide an overview of the discipline's various facets. We distinguish between *developmental models* and *developmental metrics* and relate each to different types of quality-control devices. In our view, models are either *explanatory* or *descriptive*, and their quality is evaluated in terms of specific types of disciplinary discourse. Metrics are either *calibrated measures* or *soft measures*, and their quality is evaluated in terms of specific psychometric parameters. Following a discussion on how developmentalists make metrics, and on a variety of metrics that have been made, we discuss the two key psychometric quality-control parameters, *validity* and *reliability*. This sets the stage for a limited and exploratory literature review concerning the quality of a set of existing metrics. We reveal a conspicuous lack of psychometric rigor on the part of some of the most popular developmental approaches and invite remedies for this situation.

Keywords: developmental assessment, developmental psychology, epistemology, meta-theory, psychological technologies, psychometrics, quality control, usable knowledge.

Introduction: Concern Over the Measuring of Psychological Phenomena

The marketplace of ideas is rife with various instruments, measures, and techniques designed to gear into the psychological lives of individuals and the lifeworlds of communities—from Spiral Dynamics to Requisite Organization, Integral Life Practice, and a multitude of leadership initiatives. But what are the standards that should determine which of these psychological technologies should go to market? And what of the psychological technologies to come? What are the quality-control parameters for this kind of knowledge production? These are the concerns that guide this paper. We provide an overview of the discipline's various facets in an attempt to clarify our call for quality control efforts, and guide their execution. This is the second of two papers about exercising quality control over the production of *usable knowledge* in the discipline (see Stein, 2008a). We are not the only ones worried about the market cultures consuming psychological technologies manufactured by the discipline (see Ross, 2008) and we hope to contribute to the conversation by introducing key theoretical and methodological distinctions.

More specifically, the heart of the concerns we raise here are *psychometrical*. Psychometrics is a branch of metrology, which is the transdisciplinary study of metrics. As the term implies, psychometricians are specifically concerned about measuring psychological phenomena. As explained below, developmental psychologists build *models* and *metrics*, seeking both to understand the development of psychological phenomena, and to measure them. Developmental

metrics—often called scoring systems—deserve attention because they are situated near the heart of the discipline, and radically shape both research and practice.

Below, we provide an overview of developmental psychology as a discipline, presenting a map representing its different important aspects. We then discuss this map, paying particular attention to methods for building developmental metrics. This sets the stage for a limited, exploratory literature review covering a set of metrics currently in use. According to our review, the literature reveals a conspicuous lack of psychometric rigor on the part of some developmental approaches. Generally, this paper is an invitation to the community to remedy this situation, and begin more a serious discourse about quality control in the field.

Models and Metrics in Developmental Psychology

To start, we will situate some of our key terms. Table 1 below displays four example disciplines, along with their *metrics*, *models*, and *technologies*. The distinctions we suggest—e.g., between *metrics* and *models*—are discipline-general distinctions. In other words, they are meta-theoretical distinctions, insofar as they *oversee* and *explicate* discipline-specific distinctions. Thus our concerns about *technologies* are also meta-theoretical. The question of how general and useful these distinctions are is not a question we will address in the context of this paper. Do all sciences have some variety of metric, model, and technology? It seems so from what we can see, but this paper will focus solely on developmental psychology. These distinctions are certainly relevant to developmental psychology and the issues immediately at hand. We discuss how to use these distinctions—and a host of others—to institute quality control standards for the production of usable knowledge in the field.

Table 1. Knowledge Production in Selected Disciplines

Discipline	Metrics	Models	Examples of Technologies
Developmental psychology	Psychometrically calibrated	Disciplined discourse about developmental *models*	*Psychological technologies* for education
Medicine	Biometrically calibrated	Disciplined discourse about *models* of disease	*Medical technologies* for healing
Sociology	Sociometriclly calibrated	Disciplined discourse about *models* of social life	*Social technologies* for group transformation
Economics	Econometrically calibrated	Disciplined discourse about economic *models*	*Financial technologies* for market amelioration.

Figure 1 zooms in on the first row of Table 1 to reveal more detail about developmental psychology. As discussed below, *metrics* and *models* are further characterized as *representational devices*, both with subtypes. Figure 1 relates metrics and models to their respective quality-control devices. *Technologies* are not represented because they require their own separate map; an interesting map that nonetheless will not be rendered here. Importantly,

according to our view, in developmental psychology *metrics* and *models* set the conditions for the possibility for *psychological technologies*. The two-way arrows—both horizontal and vertical—signify the complex inter-animations and relations between the various areas differentiated in the map. Metrics, models, and their respective quality control devices inter-participate in dynamic ways, especially during the generation of usable knowledge and the building of psychological technologies. However, these complexities are beyond the scope of this initial treatment of the issues. Nevertheless, despite neglecting some important topics, this paper should still be viewed as a critical meta-theory for guiding the production of future psychological technologies brokered by developmental psychologists.

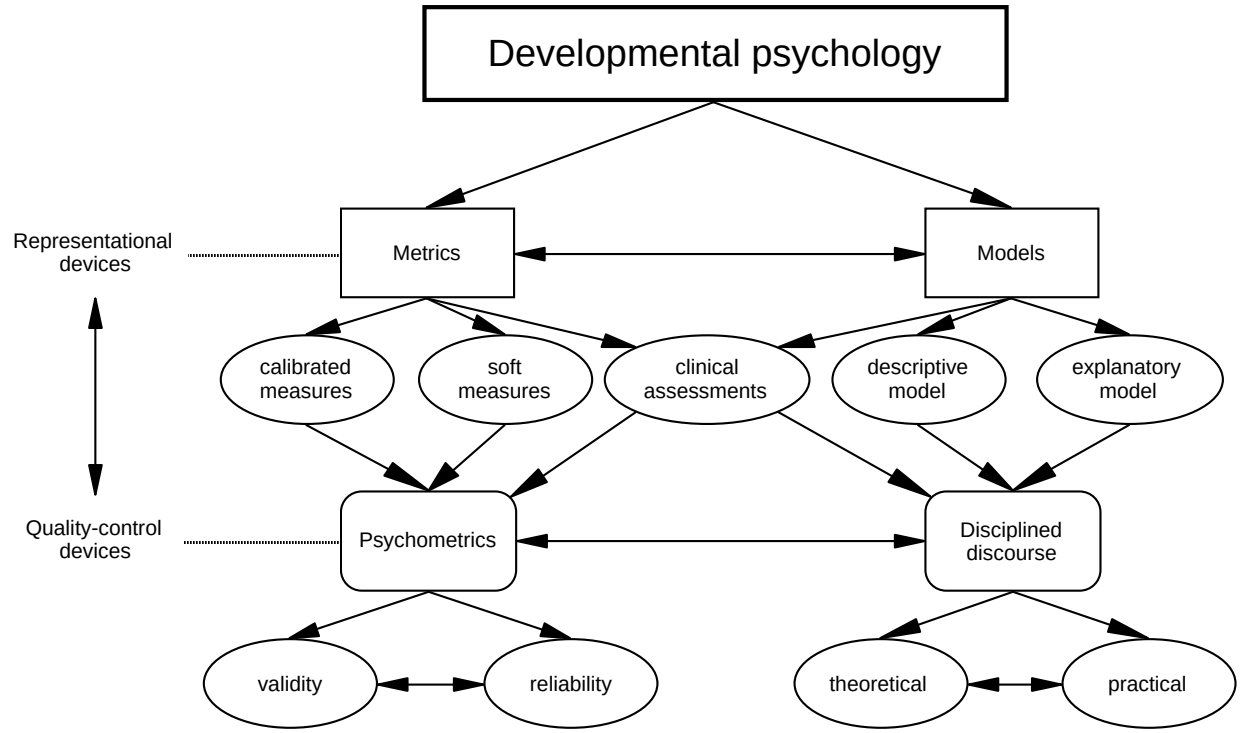

Figure 1. A Meta-theoretical Explication and Representation of Knowledge Production Processes in Developmental Psychology

The distinctions and relations displayed in the map above are based on a broad survey of comparable meta-theoretical projects concerning the epistemological workings of psychology as a discipline. Most useful in this respect are the following: Piaget's (1970a, 1970b, 1970c) work on interdisciplinarity in psychology; Habermas's (1988) work on the logic of the social sciences; Donald Campbell's (1959, 1969, 1987) work on an epistemologically relevant social psychology of the disciplines; Overton's (2006, 2007) systems-theoretic meta-theory for developmental psychology; Fischer and Bidell (2006) and Van Geert's (1994) work on model building in developmental psychology; and Dawson (Dawson-Tunik, 2004) and Jaques's (1978) work on metrics in developmental psychology. The fully explicated set of working terms garnered and synthesized from this cast is presented in Figure 2. Once again, we are simply presenting a collection of meta-theoretical distinctions topographically, as a system of related terms and categories. Figure 2 explicates Figure 1 in a slightly different representational modality.

<table>
<tr><th colspan="6">Developmental Psychology</th></tr>
<tr><td rowspan="2">Representational devices: methods, symbol systems, and propositions—built by communities of developmental psychologists—which in some way claim to be about cognitive developmental processes.</td><td colspan="2">Metrics: representational devices built and used to determine the amount of or degree to which a specific psychological trait is found to be in a sample of performances.</td><td colspan="3">Models: representational devices built and used to account for or characterize the presence of a specific psychological trait found in a sample of performances.</td></tr>
<tr><td>Calibrated measures: metrics that adhere to strict quantitative standards for determining the amount of a trait. Mainly for use in measuring individuals.</td><td>Soft measures: metrics with mainly qualitative standards for determining the amount of a trait. Mainly for use in conducting research.</td><td>Clinical assessments: judgments and diagnoses rendered by professionals, informed by models and based on clinical experience.</td><td>Explanatory models: models that posit mechanisms or processes to account for the occurrence of the trait.</td><td>Descriptive models: models offering idealized or exemplary characterizations of the trait.</td></tr>
<tr><td rowspan="2">Quality-control devices: methods created by developmental psychologists for selecting between different types of representational devices, e.g., ranking them in terms of their validity.</td><td colspan="2">Psychometrics: research and theory concerning the creation and maintenance of quality-control standards for metrics.</td><td colspan="3">Disciplinary Discourse: critical exchange of evidence and argument concerning the creating and maintenance of quality-control standards for models.</td></tr>
<tr><td>Validity: the focus of psychometric studies about the types of reasonable inferences and claims we can make in light of the results yielded by a metric.</td><td>Reliability: the focus of psychometric studies about the performance-quality of metrics, specifically about the degree and type of their error-proneness.</td><td>Theoretical: disciplinary discourse about the truth, coherence, and reasonableness of models.</td><td>Practical: disciplinary discourse about the ethics, efficacy, and implications of putting models to use.</td></tr>
</table>

Figure 2. Terms and Definitions about Knowledge Production Processes in Developmental Psychology

Figures 1 and 2 require elaboration. Examples will be useful in clarifying some of the key distinctions. A good place to begin is the work of Erik Erikson (1980), who offers perhaps the most widely known instance of a developmental model *not* tied to a metric. That is, Erikson's *model*—which describes and explains the life course in terms of eight stages of increasing maturity—has no accompanying *metric*. Such a metric would be an explicit procedure for determining the degree of maturity evidenced in an individual's language or actions. So while Erikson offers *descriptions* and *explanations* of development—based on clinical experience, and useful in honing *clinical assessments*—he offers no explicit tools for making measurements or assessments of development. Models inform our understanding insofar as they characterize and

account for things, but they don't interface with and disclose the world in the same way as metrics do.

Kohlberg (1981, 1984), on the other hand, had both a *metric* and a *model*. Following Baldwin, Mead, Dewey, and Piaget, Kohlberg offered a model of moral development wherein an individual is socialized into a world of norms and systems of mutual expectation, passing through six stages on the way to a kind of post-conventional moral autonomy. Alongside this model was Kohlberg's metric (Colby & Kohlberg, 1987a; 1987b), which would come to be known as the Standard Issue Scoring System (the SISS). The SISS was a codified, explicit procedure for *disclosing* and *determining* the degree of moral development in evidence in particular linguistic performances. Metrics gear into the properties researchers are interested in and disclose them for measurement, while models describe or explain those same properties, accounting for their being the way they are.

Importantly, both models and metrics are *representational devices*. This is a term that can be traced to Rawls (1996), who originally used it in the context of moral and political philosophy. He famously explicated the "original position"—an impartial and radically perspectival post-conventional view of social reality—in order to codify the basic tenets of a comprehensive theory of justice (Rawls, 1973). He would later retrospectively characterize the function of the "original position" by calling it a *representational device*. That is, he argued that it is best understood as a contrivance to help us *see* the world—to represent it—in terms of justice, and thus to help us hone our moral judgments about political situations. He had a broad notion of justice, and wanted to make a certain view of it readily accessible, so he built a *representational device*.

Understandably, philosophers have adopted the term for use in debates surrounding post-positivist theories of knowledge (Elgin, 2004). In this context, representational devices play a role in disciplinary knowledge production. Roughly speaking, representational devices serve as the parts of a discipline's epistemic architecture that purport to be *about* the world. That is, disciplinarians do not traffic in facts, first and foremost; they mainly traffic in representational devices, which are, more or less, a condition for the possibility of facts. From dynamic systems theorists' mathematical mock-ups to the Weberian ideal-types built by sociologists, we place representational devices between *ourselves* and *the world*, in order to arrange for a specifically desirable view.

In developmental psychology, as we have been discussing, metrics and models are the two main species of representational devices. Importantly, they come in different varieties. Models can be characterized as being more or less *descriptive* or *explanatory*. Piaget (1972) famously built both *descriptive* and *explanatory* models of cognitive development. Over the course of more than half of a century and based on observations and experiments on literally thousands of children, Piaget *described* various stage sequences (e.g., sensorimotor, pre-operational, concrete-operational, formal-operational) and suggested a set of *explanatory* constructs to account for their genesis (e.g., assimilation, accommodation, equilibration).

There are many approaches to building models in developmental psychology, but they can all be described in the terminology of a reasonably nuanced philosophy of science (Magnani &

Nersessian, 2002). Briefly, descriptive models arise from reasoning that is *inductive* (Peirce, 1986), that is, reasoning that involves generalizing from properties of a sample to properties of a population. Piaget found ways to usefully classify and describe the various types of cognitive structures he witnessed, and built a descriptive model by positing the *general applicability* of an idealized account of these cognitive structures. Explanatory models, on the other hand, arise from reasoning that is *abductive* (Peirce, 1986), that is, reasoning involving inferences to the best explanation. Piaget looked to the general principles of biological evolution and to philosophical logic in order to account for the transformations of cognitive structures he witnessed, and built an explanatory model by explicitly codifying these "guesses at the riddle."

Closely related to models and to metrics—but belonging strictly in neither category—are *clinical assessments*. This important area of developmental psychology is the province of trained professionals, from therapists to teachers and consultants. Clinical assessments are complex judgments about individuals rendered in uncertain real-life conditions, such as during the practice of psychotherapy (Sullivan, 1964). They are based on models of development endorsed by professionals and they function like metrics insofar as they classify and discriminate between persons. However, they are not as explicit and codified as true metrics; they must remain flexible and facile in the hands of the professionals who make them. Moreover, the value of different types of clinical assessments hinges in large part upon the quality of the professional practices that surround it. This also sets them apart from metrics, which are deemed valuable in terms of specific psychometric quality control parameters.

Metrics can be classified as *calibrated measures* or *soft measures*. As mentioned above, Kohlberg devised a metric for looking at moral development. Importantly, the SISS was a *soft measure* of moral development, built via mainly qualitative methods and offering a content-based means for honing in on the properties of moral reasoning indicative of its developmental sophistication. As such, it was mainly built for research purposes, and *not* for measuring individuals. Loevinger (1976) aimed to devise a metric for looking at ego-development. Specifically, she was looking to build a *calibrated measure.* By merging quantitative methods with qualitative ones and adhering to strict psychometric parameters, she constructed a *scale* of sentence-stem response-types. Although it is interesting to note that as well calibrated as Loevinger's metric was, she clearly stated that it was *not* to be used for rendering measurements of individuals (Loevinger, 1979; 1993).

On the Construction of Metrics in Developmental Psychology

The distinction between *calibrated measures* and *soft measures* is more difficult to understand than the difference between *explanations* and *descriptions*. This is because it requires some knowledge about how metrics are built in developmental psychology, and few know what goes on behind the scenes where developmentalists make their metrics. Therefore, before moving on to discuss the *quality-control devices* that should regulate the construction of models and metrics, and carrying out a limited literature review, it is worth explaining how researchers build developmental metrics. But to talk about developmental metrics requires some knowledge of *metrology*—the meta-disciplinary study of metrics.

Many who do theoretical work in *metrology* (e.g. Peirce, 2000; Piaget, 1972; Husserl, 1970) maintain that we can trace physical metrics like rulers and scales—and all the science that depends on them—back to the reliable capabilities of our organism and the concomitant sensorimotor practices that attune us to invariance in the world. These basic skills of being-in-the-world congeal into everyday practices that we take for granted (i.e., they form a part of the lifeworld). It is in the course of working to clarify this background of informal practices that we begin to build intersubjectively codified and calibrated metrics.

According to this view of metric making, the invariant psychophysical structures of our eyes account for the reliable differential responsiveness that justifies their use in gauging distances, a use eventually honed and codified during the collective construction of our metrics for length. However, the reliable differential responsiveness we take for granted when we interact with *things* differs from what it is during our interactions with *people* (Baldwin, 1906; Habermas, 1988; Wilber, 1995). Efforts at explicating deep-seated invariance in the social lifeworld have been dubbed *rational reconstructions* (Habermas, 1988, 1990). So when we set out to build metrics of psychological phenomena, it is the "invariant [psycho-social] properties and constitutive rules of the primary lifeworld…that can be made explicit [i.e. be rationally reconstructed] as a system of reference [allowing] for the transformation of communicative experience into measured data" (Habermas, 1988, p. 100).

Along these lines, we propose that developmental metrics should be understood as rational reconstructions of a kind of deep-seated intuitive knowledge that is *always already* a part of the network of practices and beliefs that constitute the lifeworld. Just as we unreflectively wield an intuitive knowledge of distance (based upon certain invariant relations between organism and environment) we also unreflectively wield an intuitive knowledge of development (based upon certain invariant psycho-social structures). Elsewhere, Stein (2008b) has marshaled suggestive empirical and theoretical arguments for this understanding of metric making in developmental psychology. This insight is important and ultimately suggests that developmental metrics are simply attempts to improve upon the ways we have always already made developmental judgments of each other and ourselves.

Again, the key to making a metric in developmental psychology is, as Habermas aptly described above, "the transformation of communicative experience into measured data." Specifically, we are suggesting that metrics in developmental psychology are built by rationally reconstructing—explicating and codifying—certain aspects of a preexisting reliable differential responsiveness, which we ordinarily use to make informal and intuitive developmental judgments, for example, for purposes of communication or education. Normally socialized adults automatically change their speech, action, and expectations when interacting with a child. They also typically change these over the course of an interaction with a stranger, as they gather a sense of where their interlocutor is coming from. Building a metric entails moving from this kind of everyday interpretation of language towards more systematic modes of differentiating between different types of linguistic performances.[1] To foreshadow: inter-subjectively codified modes of systematic differential classification lend themselves to techniques for *calibration,* and

[1] Importantly, with the exception of a few experimental paradigms (e.g., Piagetian balance beam tasks) all developmental metrics entail the interpretation of linguistic performances. This point is often overlooked.

the degree and type of calibration determines a metric's status as either a *soft measure* or a *calibrated measure*.

The history of cognitive developmental research has been, in part, a history of techniques for the systematic classification of various performances and behaviors. It was Piaget who first pioneered the use of clinical or qualitative interview techniques for cognitive developmental research. In a series of now classic studies, Piaget (1926, 1932) demonstrated how to look for and discover the structural properties of linguistic performances that are indicative of development. As scholars have noted (Mayer, 2005), the radical innovation here was the careful and systematic classification of performances in terms of their developmental level. Piaget, who was trained as a biologist, was out to differentiate between different species of thought and action, so he built a hierarchical taxonomy of cognitive-structure types for classificatory purposes. Along with his collaborators, he codified methods for systematic differential classification that geared into specific properties of performances, properties that were officially christened as *indexes of development*. The first developmental metrics were built.

In the wake of Piaget, different developmental researchers have offered a variety of hierarchical taxonomies and related *indexes of development* for use in the systematic differential classification of performances. However, the general approach to metric making has been the same. As noted by various researchers (e.g., Colby & Kohlberg 1987a; Cook-Greuter, 1999; Jaques & Carson, 1994; Loevinger, 1976; Piaget, 1926), the codification of a metric begins with a research team issuing relatively informal judgments about how the various performances in a dataset should be organized developmentally. That is, they argue about why one performance is more developed than another, suggesting various properties of the performances that should function as indexes of development. Then through iterative procedures for garnering intersubjective agreement, an explicit hierarchical taxonomy emerges and specific properties of the linguistic performances are promoted to the status of being indexes of development. That is, developmental metrics are built to function as *representational devices*; they are created to help us *see* development by privileging specific properties of linguistic performances that have proven useful for the purpose of reliable differential developmental classification.

For example, Kohlberg (Colby & Kohlberg, 1987a; 1987b) and others (Armon 1984; Fowler, 1981; Kitchener & King, 1981) built their metrics by privileging certain types of conceptual content—conferring them a status as indexes of development.[2] These *content-based developmental metrics* facilitate differential developmental classifications by explicating and codifying the types of ideas and concepts that appear at different levels (e.g., in Kohlberg's SISS certain specific notions about "the golden rule" appear at stage 3). Thus, researchers analyze linguistic performances by looking for the presence of specific conceptual content, which is taken to be indicative of developmental level. Decisions about which types of content are indicative of which levels are made in the process of codifying the metric. As explained above, standard practice has been to *bootstrap* the codification of indexical content from large longitudinal data sets, organized via the collective judgment of a research team and then refined

[2] This characterization of Kohlbergian metrics as *content based* is really just a claim about their *scoring manuals*. There is a great deal to be said for the domain specific *structural* indexes discussed in this tradition (e.g. socio-moral perspective taking), but the scoring manuals are mainly codified inventories of level-specific content.

into an inventory of level-specific conceptual content via iterative procedures for garnering inter-subjective agreement. This yields a *soft measure* that can be used for research purposes.

Metrics based on the use of *sentence-completion methods* (Cook-Greuter, 1999; Loevinger, 1976) are almost exactly the same, except instead of developmentally classifying qualitative interview performances, they are used to classify responses to sentence-stems. As explained in detail by Loevinger (1976), the initial steps toward making her metric involved a group of researchers creating a hierarchical taxonomy of sentence-stem response-types by relying on their collective informal judgments. This initial intuitive developmental classification of response-types was transformed via iterative procedures for garnering intersubjective agreement into an explicit system for the differential developmental classification of sentence-stem responses. Then Loevinger carried out quantitative procedures for determining the internal consistency and interval spacing of the response-type categories, because, as mentioned above, she was out to build a *calibrated measure*.

A third family of developmental metrics has been developed alongside content-based and sentence-completion methods. This family of metrics was built in a slightly different way, focusing on the *deep structures* of linguistic performances. These metrics facilitate the differential developmental classification of performances without bootstrapping from conceptual content or an inventory of sentence-stem response-types. Instead, these metrics facilitate developmental analyzes that focus on ostensibly universal properties indicative of development, which have been distilled in the wake of empirical research and model building.

Kurt Fischer (1980) first explicitly characterized and gave primacy to the basic developmental process of *hierarchical integration*, which had been an important but relatively implicit construct in developmental theorizing for nearly a century. This important advance in modeling development cleared the way for new analytic techniques useful in describing developmental pathways and constructing hierarchal taxonomies for the developmental classification of performances. These new analytical techniques dubbed *structural* properties as indexes of development: for example, they privileged attention toward properties like differentiation/integration, concreteness/abstractness, simplicity/complexity. This approach makes it possible to build many domain-specific hierarchical taxonomies using one general procedure for differential developmental classification. As explained elsewhere, using Wilber's terms (Stein & Heikkinen, 2008), privileging structural indexes of development allows us to measure and assess development in many lines using one metric. Dawson, Commons, Wilson, and Fischer (Dawson-Tunik, Commons, Wilson, Fischer, 2005) employed a variety of sophisticated quantitative techniques in order to refine an elegant set of *deep-structural properties indicative of development*.[3] The General Stage Scoring System (the GSSS has more

[3] Jaques and colleagues devised a comparable structure-based developmental metric, utilizing unique metric calibration techniques (Jaques, 1978) and a profound set of theoretical (Jaques 1970; 1976) and meta-theoretical accouterments (Jaques, 2002). This important strand of developmental research, theory, and practice has received only a footnote here because we are focusing on—for clarity's sake—the Neo-Piagetian tradition of structural metrics. However, in later work Jaques and colleagues (Jaques & Cason, 1994) have recognized their relations to this tradition, citing Fischer and Commons and drawing parallels between their respective metrics. (Cont'd next page).

recently been named The Hierarchical Complexity Scoring System, the HCSS; but see: Commons & Richards, 1984) and the Model of Hierarchical Complexity (Commons, Trudeau, Stein, Richards, & Krause, 1998) were devised. The construction of the *Lectical Assessment System* (the LAS: Dawson, 2003, 2002, 2008) took off from this tradition of deep-structural analysis.

Importantly, steps toward specific quantitative refinements of an otherwise qualitatively built metric—the steps that Loevinger and Dawson took, which Kohlberg did not—are what separates *calibrated measures* from *soft measures*. This excursus has set the stage for the discussion of *psychometrics* below, where we discuss in some detail what it means to calibrate a metric. However, at this point it should be said that we are definitively *not* suggesting that *calibrated measures* are, in principle, better than soft measures (or that these metrics are in principle better than *clinical assessments*). Depending on the situation and the metrics in question, a calibrated measure can be misleading where a soft measure is revelatory, and vice-versa. Moreover, a metric that was once calibrated can become soft when quantitative evidence wanes (e.g., Cook-Greuter, 1999). And a metric that was once soft can become calibrated if the accumulation of quantitative evidence waxes (e.g., Dawson-Tunik et al, 2005).

Nevertheless, calibrated measures and soft measures are different, and should be treated differently. Crucially, the inferences we are respectively entitled to make in light of them are not the same. For example, as mentioned above, if we want to measure *individual performances* we need to use a certain type of calibrated measure capable of making *reliable* fine-grained distinctions. If our categories are too loose and our expected measurement errors are too great, then we cannot reasonably say one performance is really more developed than another. This leads us into discussions about the *validity* and *reliability* of different types of metrics, which broaches the topic of *quality-control devices* more generally and the kinds of procedures that *ought* to be regulating the production of usable knowledge in developmental psychology.

Quality-Control Parameters: Building Useful Languages of Evaluation

We have seen how metrics are made and we have talked about models. Now we get to the heart of the matter, which is concerns about how we can make evaluations concerning the worth of specific metrics and models. Discussions about which metric or model is better than which other require decisions about how to determine their respective values. We are proposing a specific *modus operandi* for determining these values. We propose discussing the worth of metrics and models by adopting specific evaluative vocabularies—*languages of evaluation*[4]—that focus on the most important qualities that bear on the worth of both metrics and models. We

It is also worth noting here the influential model and metric devised by Kegan (1982; 1994), which deserves to be classed as a structure-based metric. However, while Kegan can be thought of as in the lineage of Piaget, he has never brought his work in line with the comparable Neo-Piagetain structure-based metrics discussed above.

[4] Stein (2007) has overseen the construction of comparable evaluative vocabularies elsewhere. The idea and phraseology can be traced in part to Taylor (1985). Although, Emerson and Marcus Aurelius toyed with the idea that to be human is to create and traffic in "vocabularies of praise."

dub these kinds of lexicons *quality control devices*. Inquiry-oriented communities of research and practice should, and often do, codify standards of quality for the representational devices they endorse and use. We are suggesting a way forward in this regard for certain areas of developmental psychology.

Figure 1 above specifies the key quality parameters that should guide the evaluative vocabularies we choose to deploy when considering different metrics and models. The distinction between *theoretical* and *practical* disciplined discourses is offered in light of a variety of comparable distinctions commonly made in epistemology and the philosophy of science. Habermas (1984, 1987) offers perhaps the most convincing and high-profile recent account, where he marshals his formal pragmatic analytical methods to justify a specific taxonomy of discourse types. Habermas includes a distinction between theoretical discourse (about the objective world; roughly cognitive-scientific) and practical discourse (about the social world; roughly moral-pragmatic). Importantly, just as descriptive and explanatory models come in many forms, there are many ways to tackle practical and theoretical disciplinary discourses.

For example, Habermas (1993) further specifies the topography of *practical discourse* by differentiating *pragmatic*, *moral* and *ethical* sub-discourses. These three focuses of practical discourse concern how we *ought* to be using our models: *pragmatic* concerns are about model's problem-focused efficacy; *ethical* concerns are about the value of the problems the model can be used to solve; *moral* concerns are about the social constraints on the proper use of the model. The point is not that Habermas' is the only evaluative vocabulary on offer, but rather that practical discourse represents a key area of concern—a key quality control parameter—that should orient the construction of some kind of related language of evaluation.

Importantly, Peirce (1931) offers his own comparable differentiations, likewise clearly partitioning practical discourse from theoretical discourse in his broad taxonomy of inquiry-types. He further specifies the topography of *theoretical discourse* by differentiating *utility*, *security* and *uberty* as the topics of specific sub-discourses (Peirce, 1998). These three focuses of theoretical discourse are roughly what are typically considered a model's scientific acumen: a model's *utility* is judged in terms of its functional-fit as a means-to-an-end; a model's *security* is a matter of the coherence and type of evidence in its support; a model's *uberty* concerns its suggestiveness and world-disclosing power. Again, the point is not that we must adopt Peirce's vocabulary, but that we must have *some* shared vocabulary to deploy when discussing the theoretical value of a model.

Thus, the goal of introducing these terms is to beef up the language of evaluation we bring to bear when we talk about the worth of various models produced by developmental psychologists. For example, take Wilber's (1999) early model of human development, as offered in the *Atman Project*. If we organize a discussion of its value around the terms just introduced, we can generate a rough profile of the model's practical and theoretical worth. Regarding its *theoretical* aspects, we think the *Atman Project* warrants valuing because of the model's *uberty*. That is, in our view, the model remains one of the only full-spectrum accounts of human development, and as such remains impressively suggestive as a rough overview of human potentials. This leads the model to have some minimal *utility* as a heuristic for various purposes, for example, alterations of one's action-orienting self-understanding. We do not think—and neither does Wilber—that the

model boasts a great deal of *security*, however. That is, many of the particular claims are in need of revision in light of existing evidence, and many of the explanations and descriptions offered are only roughly correct. Regarding its *practical* aspects, we think the *Atman Project* is valuable. Its stated use as an overview of human potentials positions the model admirably in *ethical* and *moral* discourses. Who can object to what is ultimately a positive psychology? But the devil is in the *pragmatic* considerations; specifically, how can a model that is so general address any but the most abstract problems?

Now, the point here is not to merely pass judgment on models from the past. Our goal is to demonstrate how the evaluative vocabulary we suggest might orient our model-condoning discourses. At issue is not the evaluative judgments just rendered, but rather the terms we use to render these kinds of judgments about developmental models at all. At issue is how we arrange our language of evaluation and thus set the parameters of our concerns about quality control. Of course, we have no illusions about the selective and parochial nature of the specific vocabularies deployed above. There is a great deal to say when evaluating models, much more than we cover in this brief introduction. Questions of how and why a model is *true enough* (Elgin, 2004) are as complex and heated as questions about how it *ought* to be used (Habermas, 1971).

Psychometric Quality-control Parameters: Validity and Reliability

It should be clear from the symmetry of Figure 1 that the same idea holds when we switch to discuss psychometrics. The broad distinction between *reliability* and *validity* is canonical. But while there is agreement about this general differentiation of psychometric concerns, there are many different ways to unpack validity and reliability in more specific terms. Of particular use in this regard are the *Standards for Educational and Psychological Testing*, a document that has been recursively revised for over three decades by the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME) (1999). The *Standards* represent the considered judgments of literally hundreds of disciplinary experts about the kinds of quality control standards that ought to be in place for metrics of various types; they guide the discussion that follows.

Concerns about the *validity* of a metric are concerns about the legitimate inferences that can be made about the measurements it yields. There are many different kinds of validity including but not limited to construct validity, content validity, ecological validity, and external validity.[5] Depending on the metric in question and the uses to which it is put, different aspects of validity become relevant. But all types of validity concern the question: *does the evidence suggest that the metric is measuring what it claims to be measuring?* It is important that the results yielded by a metric be valid—that we can be sure they are an index of what we think they are—so that we can make reasonable and responsible inferences based on them.

For example, it is generally agreed that *construct validity* is the most important dimension of psychometric concern (Messick, 1980). Evaluating a metric in terms of its construct validity entails looking into various types of evidence concerning how the results yielded by the metric

[5] Definitions for all these terms can be found in any standard reference (e.g., Colman, 2001; AERA et al, 1999).

conform to expectations about the postulated construct or trait it is meant to be measuring. That is, if the metric is meant to measure unidimensional wave-like growth—as certain neo-Piagetain metrics are, e.g., Dawson-Tunik et al. (2005)—then the result yielded by the metric should display a certain type of specific patterning. A metric with high construct validly allows us to make sophisticated inferences, guided by justified beliefs that we are, in fact, measuring what we claim to be. Of course, depending on what the metric claims to measure and on the uses to which it is put, different degrees and type of validity apply (AERA et al, 1999). But in general, we need information about the validity of a metric before we can responsibly do anything with it at all (Messick, 1980).

Being concerned about *reliability*, on the other hand, entails looking into how well a metric performs its function as a measure. Put another way, this means considering the kinds of errors that are likely to systematically affect the results yielded. There are many different dimensions of reliability, including but not limited to inter-rater reliability, internal consistency reliability, and test-retest reliability.[6] But all forms of reliability concern issues of *measurement error*. It is important that we know the reliability of our metrics—that we know their degree and type of accuracy—so we can use them appropriately.

For example, *internal consistency reliability* is a quantitative index of how the items or levels of a metric function in relation to one another. As explained above, every metric is a system of categories for use in the classification of performances. Insight into the internal consistency of a metric gives us a sense of how much noise—how much *measurement error*—surrounds each category. Importantly, if the measurement error is too great—if the categories are too fuzzy or noisy—then few categories can be reliably distinguished from one another, and this reduces the accuracy of the metric.

Inter-rater reliability also bears on issues of measurement error. This form of reliability concerns the noise surrounding the metric as it is put to use by different human raters—i.e., different researchers using the metric. That is: how much consensus is there between two or more raters when they award a particular performance a particular level score? Low inter-rater reliability is an index that the results yielded by the metric are noisy; a certain performances is relatively likely to be given two different scores by two different raters. High inter-rater reliability is an index of the opposite; the performances are likely to be given the same scores.

Noisy metrics with reasonable measurement error are *soft measures*, and can be very useful for research purposes that place less demand for strict precision in differential classification. Metrics with very little measurement error are *calibrated measures*, and in some cases can be used to reliably measure individual performances because they offer a precise means of differential classification. However, as with validity, depending on the metric in question and the use to which it is put, different degrees and types of reliability become relevant. But in general, we need information about the reliability of a metric before we can responsibly make use of the results it yields.

[6] Definitions for all these terms can be found in any standard reference (e.g., Colman, 2001; AERA et al, 1999).

Limited and Exploratory Literature Review

This brief overview of psychometric quality control parameters is meant to serve only as an introduction to the kind of language of evaluation we ought to employ to distinguish between the value of different metrics. Importantly, as mentioned above, metrics claming to be *soft measures* should be held to different standards from those that claim to be *calibrated measures*. That is, *soft measures* have different *reliability and validity profiles* than *calibrated measures*. But because this kind of information is essential to how we can appropriately use each measure, we are suggesting that we ought to issue some kind of reliability and validity profile for *every* metric in circulation. This would amount to adopting an explicit and shared multidimensional psychometric evaluation matrix, representing the quality control standards we think should regulate the production of metrics.

To this end we offer Table 2, which represents the result of a *limited and exploratory* literature review of metrics currently in play. *This is not meant to be an exhaustive survey*; instead our intention is to evoke the state of the readily available literature. First, we sought out references to validity and reliability studies in those key publications of the primary authors that we happened to have on hand. Second, we conducted searches in two databases: Google Scholar and PsycInfo. We reviewed the first 20-30 citations listed, passing over most dissertations and all conference papers. Any citation that included validity or reliability studies was then included. To give the reader an idea of the relative sizes of the available pools of literature, the number of hits from the primary search term in each database is listed in parentheses in the footnotes.

Table 2: Exploratory Review of Validity and Reliability Studies for a Set of Developmental Metrics

Metric	Source	Aspect of psychometric quality addressed	Publication type
Leadership Development Profile[7]	None found.		
Spiral Dynamics[8]	None found.		
SCTi / MAP[9]	Cook-Greuter, 1999	Content validity, inter-rater reliability	Dissertation
Requisite Organization[10]	Jaques, 1978	Construct validity	Book
	Jaques & Cason, 1994	Inter-rater reliability	Book

[7] Search terms: "leadership development profile" (hits: 35; 0), "leadership development profile" reliability, "leadership development profile" validity. Rooke and Torbert (1998) and several other Torbert works cite Loevinger's studies as confirmation of the LDP's reliability and validity. But Loevinger's studies are out of date (new raters, scoring criteria, and levels are now used; metrics need to be re-calibrated), so they are not included here.

[8] Search terms: "spiral dynamics" (hits:879; 11) "spiral dynamics" "people scan," "spiral dynamics" validity, "spiral dynamics" reliability. Includes a large number of irrelevant hits having to do with physical science.

[9] Search terms: SCTi (irrelevant results), MAP assessment (irrelevant results), "Susanne Cook-Greuter" (hits: 71; 2).

[10] Search terms: "requisite organization" (hits: 526; 4), "requisite organization" reliability, "requisite organization" validity.

Metric	Source	Aspect of psychometric quality addressed	Publication type
Subject Object Interview[11]	Lahey, Souvaine, Kegan, Goodman, and Felix, 1988	Inter-rater reliability, test–retest reliability, construct validity	Self-published training manual
	Pratt, Diessner, Hunnsberger, Pancer, and Savoy, 1991	Inter-rater reliability, construct validity	Peer-reviewed journal
	Harris & Kuhnert, 2008	Inter-rater reliability, predictive validity	Peer-reviewed journal
	Lewis, Forsyth, Sweeney, Bartone, Bulls, and Snook, 2005	Inter-rater reliability, predictive validity	Peer-reviewed journal
Hierarchical Complexity Scoring System (General Stage Scoring System)[12]	Dawson 2003	Congruent validity, internal consistency, inter-rater reliability	Peer-reviewed journal
	Dawson, 2004	Congruent validity	Peer-reviewed journal
	Dawson & Gabrielian, 2003	Congruent validity	Peer-reviewed journal
	Dawson, Xie, & Wilson, 2003	Congruent validity, internal consistency, inter-rater reliability	Peer-reviewed journal
	Dawson-Tunik, Commons, Wilson, & Fischer, 2005	Construct validity, internal consistency, Inter-rater reliability, evidence of fine-grained interval scale	Peer-reviewed journal
	Xie & Dawson, 2006	Construct validity	Peer-reviewed journal
Lectical Assessment System[13]	Dawson,* 2000	Congruent validity, internal consistency, inter-rater reliability	Peer-reviewed journal
	Dawson,* 2002	Congruent validity, inter-rater reliability	Peer-reviewed journal
	Dawson, 2006	Construct validity, evidence of fine-grained interval scale	Book chapter
	Dawson-Tunik, 2004	Construct validity, internal consistency, evidence of fine-grained interval scale	Peer-reviewed journal

Importantly, Table 2 is not meant to be an exhaustive literature review. We know we have missed some studies and left a few metrics out. But the table is offered as a kind of proof-of-

[11] Search terms used: "Subject Object Interview" (hits: 105; 30), "Subject Object Interview" validity, "Subject Object Interview" reliability, "Subject Object Interview" interrater reliability, "constructive development theory"

[12] Search terms used: "hierarchical complexity scoring system" (hits: 49; 9). System had a name change from GSSS to HCSS.

[13] Search terms used: “Lectical Assessment System” (hits:12; 0). Studies with an asterisk are those conducted on an early version of the LAS, temporally named the HCSS. However, despite all name changes, the LAS is significantly different from the HCSS previously known as the GSSS.

concept, an example of the way our concerns about quality control interface with knowledge production in the discipline. In this respect, the table is an invitation to the community; we would like to see comparable tables built by others, especially by the actual researchers concerned. We are just trying to kick-start a concerted discourse about quality control in the field.

From what we could find, there is no evidence that the Online PeopleScan by Spiral Dynamics Integral offers reliable or valid measurements of anything. In this preliminary review, it also appears that contemporary uses of sentence completion methods to measure ego-development are not justified by peer-reviewed validity and reliability studies, and what studies there are do not address the most important aspects of psychometric concern (i.e., construct validity and internal consistency). In fact, in light of this exploratory literature review, it appears that the LAS and HCSS are the only metrics that have been calibrated using quantitative indexes of internal consistency. This means that the LAS and HCSS are the only ones that can be validly and reliably used to assess individuals; they are the only *calibrated metrics* in the set, the rest are *soft measures* that should only be used for research purposes. If we have missed publications (or unpublished white papers) we apologize, but offer this table as an invitation for greater transparency and integrity amongst researchers. Let us work together to fill out and expand the table, making clear to each other and to all those affected exactly what we are doing.

If we want to see an integral and developmental worldview gain a real institutional foothold—radically reforming business, government, education, therapy, and our own sense of human potentials—we need to get serious about our quality control standards. How can we expect to change *status quo* means of human resource management, educational testing, and mental health care when their knowledge production processes are more rigorous and professionalized than ours? Of course, there are more than merely pragmatic concerns on the table. As explained elsewhere (Stein, 2008a), there are important ethical issues in play. Concerns about psychometrics dovetail with concerns about *justice* insofar as our measures affect the lives of individuals and structure the shape of communities (Fisher, 2004; 2005). The call for quality control issued here is not a merely academic exercise.

Conclusions: Toward an Integral and Problem Focused Metrological Pluralism

In this paper we have offered an overview of knowledge production processes in developmental psychology, with an eye toward instituting quality control parameters for the creation of usable knowledge in the discipline. Clearly, the cursory treatment offered here requires further elaboration. In particular, the complex relations between metrics, models, and their respective quality control devices has been left for another day. Likewise, a very important discussion needs to take place concerning the transition from theory and research to practice, that is, about the methods we use in the creation and dissemination of psychological technologies. But we have not been seeking to offer *the* definitive account. We are really just pointing to the kind of reflective explication and survey that should be done by those concerned about the production of this kind of usable knowledge.

We have focused on metric-making in particular because it seems to be the least understood aspect of the discipline. We traffic in models all the time, and are more or less used to judging

their worth (even if we have no shared language of evaluation). But metrics play a very specific role in developmental psychology, as embodiments of usable knowledge. As Table 2 begins to demonstrate, there are many metrics currently in play and they are not all equally valid and reliable. The kinds of validity and reliability profiles we are suggesting would go a long way toward raising the level of our discourse about developmental metrics. Importantly, a sophisticated discourse about metrics would *not* necessarily result in a homogenization of metric types or a marginalization of metrics with certain profiles. In fact, we firmly believe that such a discourse about metrics would move us toward an *integral and problem focused metrological pluralism*.

The ideal of an *integral and problem focused metrological pluralism* suggests that we reflect on the way we use developmental metrics, in order to make principled decisions about which types of metrics are appropriate for which purposes. In order to do this we need to build validity and reliability profiles for existing metrics and begin to dialogue about the implications of what we have to work with. This rigorous reflection on metrics should be supplemented with reflections on the problem-spaces we face. What are we trying to do with the metrics we have? For example, it is likely that clinicians and coaches need different metrics than researchers, while educators require measures that differ from those used by organizational consultants.

This ideal would also suggest that we need to study the efficacy of our efforts, and feed this data back into our reflections about which metrics are good for which purposes (Stein, 2008a). Currently, our success stories are anecdotal, so our enthusiasms seem a bit naïve. As the limited and exploratory literature review above suggests, we do not really know the nature of the instruments we wield. We need real *clarity* on a variety of issues—including not only psychometric ones, but also ethical, political, and pedagogical ones—before we can reasonably and responsibly move forward. To our minds, this clarity is a function of collaboration.

An *integral and problem focused metrological pluralism* ultimately requires open collaborative efforts toward building usable knowledge. A first step would be expanding and filling out Table 2 above, that is, building something like an information clearinghouse for developmental approaches. This kind of transparency among researchers, and between researchers and the public, could catalyze the growth of the discipline, moving us away from a set of cloistered consultancies and toward a network of collaboration, knowledge sharing, and rigorous debate. Thus, we echo the call for something like an *Institute for Applied Developmental Theory*, as articulated by Ross (2008) on prior pages of this journal. More specifically, this paper is offered as an invitation to those who might want to shape the future of the field and—with the support of *Integral Review* editorial staff—invite responses from the community.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Armon, C. (1984). *Ideals of the good life: Evaluative reasoning in children and adults.* Unpublished Doctoral dissertation, Harvard University, Boston.

Baldwin, J. M. (1906). *Thought and things: A study in the development of meaning and thought or genetic logic* (Vol. 1-4). New York: Macmillan Co.

Campbell, D. T. (1959). Methodological suggestions from a comparative psychology of knowledge processes. *Inquiry, 2*, 152-182.

Campbell, D. T. (1969). Ethnocentrism of disciplines and the fish-scale model of omniscience. In *Interdisciplinary relationships in the social sciences (Sherif and Sherif, ed.)*. Chicago: Aldine.

Campbell, D. T. (1987). Evolutionary epistemology. In Radinitzky & Bartly (Eds.), *Evolutionary epistemology, theory of rationality, and the sociology of knowledge*. La Salle, IL: Open Court.

Colby, A., & Kohlberg, L. (1987a). *The measurement of moral judgment, vol. 1: Theoretical foundations and research validation*. New York: Cambridge University Press.

Colby, A., & Kohlberg, L. (1987b). *The measurement of moral judgment, vol. 2: Standard issue scoring manual*. New York: Cambridge University Press.

Colman, A. M. (2001). *Dictionary of psychology*. New York: Oxford University Press.

Commons, M. L., & Richards, F. A. (1984). Applying the general stage model. In M. Commons, F. A. Richards & C. Armon (Eds.), *Beyond formal operations* (pp. 147-157). New York: Praeger.

Commons, M. L., Trudeau, E. J., Stein, S. A., Richards, F. A., & Krause, S. R. (1998). Hierarchical complexity of tasks shows the existence of developmental stages. *Developmental Review, 18,* 238-278.

Cook-Greuter, S.R. (1999). *Postautonomous ego development: A study of its nature and development*. Doctoral Dissertation. Harvard Graduate School of Education.

Davidson, D. (2001). *Subjective, intersubjective, objective*: Cambridge University Press.

Dawson, T.L. (2000). Moral reasoning and evaluative reasoning about the good life. *Journal of Applied Measurement, 1*, 372-397.

Dawson, T.L. (2002). A comparison of three developmental stage scoring systems. *Journal of Applied Measurement, 3(2),* 146-189.

Dawson, T.L. (2003). A stage is a stage is a stage: A direct comparison of two scoring systems. *Journal of Genetic Psychology, 164,* 335-364.

Dawson, T.L. (2004). Assessing intellectual development: Three approaches, one sequence. *Journal of Adult Development, 11(2*), 71-85.

Dawson, T.L. (2006). Stage-like patterns in the development of conceptions of energy. In X. Liu & W. Boone (Eds.), *Applications of Rasch measurement in science education* (pp. 111-136). Maple Grove, MN: JAM Press.

Dawson, T. L. (2008). The Lectical™ Assessment System. 1. Retrieved July, 2008, from http://www.lectica.info

Dawson, T.L., & Gabrielian, S. (2003). Developing conceptions of authority and contract across the life-span: Two perspectives. *Developmental Review, 23*, 162-218.

Dawson, T.L., Xie, Y., & Wilson, M. (2003). Domain-general and domain-specific developmental assessments: Do they measure the same thing? *Cognitive Development, 18*, 61-78.

Dawson-Tunik, T. L. (2004). "A good education is..." The development of evaluative thought across the life-span. *Genetic, Social, and General Psychology Monographs*, 130, 4-112.

Dawson-Tunik, T. L., Commons, M. L., Wilson, M., & Fischer, K. W. (2005). The shape of development. *The International Journal of Cognitive Development, 2*, 163-196.

Elgin, C. (2004). True enough. *Philosophical issues, 14*, 113-131.

Erikson, E. (1980). *Identity and the life cycle*. New York. W. W. Norton & Co, Inc.

Fischer, K. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review, 87*(6), 477-531.

Fischer, K., & Bidell, T. (2006). Dynamic development of psychological structures in action and thought. In W. Damon & L. R.M. (Eds.), *Handbook of child psychology: Theoretical models of human development* (Vol. One, pp. 1-62). New York: John Wiley & Sons.

Fisher, W.P. (2004) Meaning and method in the social sciences. Human Studies. 27, 429-454.

Fisher, W.P. (2005) Mathematics, measurement, metaphor, and metaphysics: implications for

Fowler, J. (1981). *Stages of faith: The psychology of human development and the quest for meaning*. San Francisco: Harper and Row.

Habermas, J. (1971). *Knowledge and human interests.* Boston: Beacon Press.

Habermas, J. (1984). *The theory of communicative action: Reason and the rationalization of society* (T. McCarthy, Trans. Vol. 1). Boston: Beacon Press.

Habermas, J. (1987). *The theory of communicative action: Lifeworld and system, a critique of functionalist reason* (T. McCarthy, Trans. Vol. 2). Boston: Beacon Press.

Habermas, J. (1988). *On the logic of the social sciences* (Nicholsen & Stark, Trans.). Cambridge: MIT Press.

Habermas, J. (1990). Reconstruction and interpretation in the social sciences (Nicholsen, Trans.). In *Moral consciousness and communicative action* (pp. 21-43). Cambridge, MA: MIT Press.

Habermas, J. (1993). On the pragmatic, ethical, and moral employments of practical reason. In *Justification and application* (pp. 1-19). Cambridge, MA: MIT Press.

Harris, L.S. and Kuhnert, K.W. (2008). Looking through the lens of leadership: A constructive development approach. *Leadership and Organization Development Journal, 29(1),* 47-67.

Husserl, E. (1970). *The crisis of European sciences and the transcendental phenomenology*. (Carr, Trans). Evenston, IL. Northwestern University Press.

Jaques, E. (1970). *Work, creativity, and social justice*. London: Heinemann Educational.

Jaques, E. (1976). *A general theory of bureaucracy*. London: Heinemann Educational.

Jaques, E. (1978). *Levels of abstraction in logic and human action: A theory of discontinuity in the structure of mathematical logic, psychological behaviour, and social organisation*. London: Heinemann Educational.

Jaques, E. (2002). *The life and behavior of living organisms: A general theory*. Westport, CT: Praeger.

Jaques, E., & Cason, R. (1994). *Human capability: Study of individual potential and its application*. Fall Church, VA: Cason Hall.

Kegan, R. (1982). *The evolving self: Problem and process in human development*. Cambridge, Massachusetts: Harvard University Press.

Kegan, R. (1994). *In over our heads: The mental demands of modern life*. Cambridge, MA: Harvard University Press.

Kitchener, K. S., & King, P. M. (1981). Reflective judgment: Concepts of justification and their relationship to age and education. *Journal of applied developmental psychology, 2*, 89-116.

Kohlberg, L. (1981). *The philosophy of moral development: Moral stages and the idea of justice* (Vol. 1). San Francisco: Harper & Row.

Kohlberg, L. (1984). *The psychology of moral development: The nature and validity of moral stages* (Vol. 2). San Francisco: Jossey Bass.

Lahey, L., Souvaine, E., Kegan, R, Goodman, R., & Felix, S. (1988). *A guide to the Subject-Object Interview: Its administration and interpretation*. Cambridge, MA: Harvard Graduate School of Education, Subject-Object Research Group.

Lewis, P., Forsythe, G.B., Sweeney, P., Bartone, P., Bullis, C., Snook, S. (2005). Identity development during the college years: Findings from the West Point Longitudinal Study. *Journal of College Student Development, 46(4),* 357-373.

Loevinger, J. (1976). *Ego development*. San Francisco: Jossey Bass.

Loevinger, J. (1979). Construct validity of the sentence completion test of ego development. *Applied Psychological Measurement,* 3(3), 281-311.

Loevinger, J. (1993). Ego development: Questions of method and theory. *Psychological Inquiry, 4*(1), 56-63.

Magnani, L., & Nersessian, N. J. (Eds.). (2002). *Model-based reasoning: Science, technology, values*. New York: Acdemic/Plenum Publishers.

Mayer, S. J. (2005). The early evolution of jean piaget's clinical method. *History of psychology, 8*(4), 362-382.

Messick, S. (1980) Test validity and the ethics of assessment. *American psychologist.* 35(11), 1012-1027.

Overton, W. (2006). Developmental psychology: Philosophy, concepts, and methodology. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology: Theoretical models of human development* (6 ed., pp. 18-88). New York: John Wiley & Sons.

Overton, W. F. (2007). A coherent metatheory for dynamic systems: Relational organicism-contextualism. *Human development, 50,* 154-159.

Peirce, C. S. (1984). On a new list of categories. In The Pierce Edition Project (Ed.), *Writings of Charles S. Peirce: A chronological edition* (vol. 2). Bloomington, IN: Indiana University Press.

Peirce, C. S. (1986). Deduction, induction, and hypothesis. In P. e. project (Ed.), *Writings of charles s. Peirce: A chronological edition* (Vol. 3). Bloomington, IN: Indiana University Press.

Peirce, C. S. (1998). An essay toward reasoning in security and uberty. In *The Essential Peirce (ed. Peirce edition project)* (Vol. 2). Bloomington, Indiana: Indiana University Press.

Peirce, C. S. (2000). *The writings of Charles S. Peirce: A chronological edition* (vol. 1-6). Edited by the Peirce Edition Project. Bloomington, IN: Indiana University Press.

Piaget, J. (1926). *The language and thought of the child*. London: Lund Humphries.

Piaget, J. (1932). *The moral judgment of the child*. London: Routledge and Kegan Paul.

Piaget, J. (1970a). *Main trends in psychology*. New York: Harper & Row.

Piaget, J. (1970b). *Main trends in psychology in interdisciplinary research*. New York: Harper & Row.

Piaget, J. (1970c). *The place of the sciences of man in the system of sciences*. New York: Harper & Row.

Piaget, J. (1972). *The principles of genetic epistemology* (W. Mays, Trans.). London: Routledge & Kegan Paul.

Pratt, M.W., Diessner, R., Hunsberger, B., Panser, S.M., Savoy, K. (1991). Four pathways in the analysis of adult development and aging: Comparing analyses of reasoning about personal-life dilemmas. *Psychology and Aging, 6(4),* 666-675.

Rawls, J. (1973). *A theory of justice*. London: Oxford University Press.

Rawls, J. (1996). *Political liberalism*. New York: Columbia University Press.

Rooke, D. and Torbert, W.R. (1998). Organizational transformation as a function of CEO's developmental stage. *Organization Development Journal, 16(1),* 11-28.

Ross, S.N. (2008) Using developmental theory: When not to play the telephone game. *Integral Review*. Vol. 4 No. 1.

Stein, Z & Heikkinen, K. (2008). On operationalizing aspects of altitude: an introduction to the Lectical Assessment System for Integral researchers. *Journal of Integral Theory and Practice*. Spring. Vol 3. No 1.

Stein, Z. (2007). Modeling the demands of interdisciplinarity. *Integral Review*. *4*.

Stein, Z. (2008a) Myth busting and metric making: refashioning the discourse about development. *Integral Leadership Review.* Vol. 8. No. 5.

Stein, Z. (2008b) Intuitions of altitude: researching the conditions for the possibility of developmental assessment. Paper presented at Biannual Integral Theory Conference, John F. Kennedy University. Pleasant Hill, CA.

Sullivan, H. S. (1964). *The collected works of Harry Stack Sullivan* (Vol. 1 and 2). New York: W.W. Norton & Company.

Taylor, C. (1985). *Human agency and language: Philosophical papers* (Vol. 1). Cambridge: Cambridge University Press.

Van Geert, P. (1994). *Dynamic systems of development: Change between complexity and chaos.* New York: Harvester Wheatsheaf.

Wilber, K. (1995). *Sex, ecology, spirituality - the spirit of evolution*. Boston: Shambhala Publications.

Wilber, K. (1999). The Atman Project. In *Collected works of Ken Wilber* (Vol. 2). Boston: Shambhala Publications.

Xie, Y., & Dawson, T. L. (2006). Multidimensional models in a developmental context. In M. Garner, G. Engelhard, M. Wilson & W. Fisher (Eds.), *Advances in Rasch Measurement*. JAM Press.

Zachary Stein *is a student of philosophy and cognitive development pursing a doctorate at the Harvard University Graduate School of Education. He is also the senior analyst at the Developmental Testing Service.*
Email: zas@mail.harvard.edu

Katie Heikkinen *is doctoral student at the Harvard Graduate School of Education, focusing on adult development—its assessment and its relationship to leadership outcomes. She is also on the faculty of the Integral Theory program at John F. Kennedy University, where she teaches multiple intelligences theory and cognitive development.*
Email: katie.heikkinen@gmail.com