UNIVERSITY OF CALIFORNIA

Santa Barbara

A Discovery-Based Workflow for Educational Measurement

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Education

by

Anthony Clairmont

Committee in charge:

Professor Andrew Maul, Co-Chair

Professor Jin Sook Lee, Co-Chair

Professor Mary E. Brenner

Professor Chris Newfield

December 2020

The dissertation of Anthony Clairmont is approved.

Chris Newfield

Mary E. Brenner

Jin Sook Lee, Committee Co-Chair

Andrew Maul, Committee Co-Chair

December 2020

A Discovery-Based Workflow for Educational Measurement

by

Anthony Clairmont

Acknowledgements

Vita

Anthony Clairmont

EDUCATION

2020 PhD in Education, University of California, Santa Barbara
2016 MA in Education, University of California, Santa Barbara
2012 MA in French Civilization & Culture, Middlebury College
2011 BA in Philosophy, Sewanee: the University of the South

PUBLICATIONS

2020 Clairmont, Anthony. Wolf, Melissa G., & Maul, Andrew. "The prevention and detection of deception in self-report survey data." In *Basic Elements of Survey Research in Education: Addressing the Problems Your Advisor Never Told You About,* eds. Ulemu Luhanga & Gregg Harbaugh. Charlotte, NC: Information Age Publishing.

2020 Arya, Diana, Clairmont, Anthony**,** Katz, Daniel, & Maul, Andrew. "Measuring Reading Strategy Use." *Educational Measurement* (special issue, ed. Michael Kane).

2020 Arya, Diana, Clairmont, Anthony, & Hirsch, Sarah. "Interpreting and explaining data representations: A comparison across grades 1-7." In *Approaches to Lifespan Writing Research: Steps Toward an Actionable Coherence,* eds. Ryan Dippre & Talinn Phillips. Colorado: The WAC Clearinghouse, Colorado State University Press, & the University Press of Colorado.

2019 Clairmont, Anthony, & Maul, Andrew. "A Review of *Sociocognitive Foundations of Educational Measurement* by Robert Mislevy." *Psychometrika*, 84(4), 1097-1100.

Abstract

A Discovery-Based Workflow for Educational Measurement
by
Anthony Clairmont

Educational researchers are often tasked with postulating and measuring unknown processes that account for observed outcomes. Logically, this requires inductive reasoning about which constructs are relevant to the situation we seek to understand, prior to attempts at measurement. Exploratory mixed methods designs, along with a few classic designs from psychometrics, represent the canonical approaches to addressing this research problem. However, I argue that these canonical approaches require rethinking in order to fully embrace both the logic of discovery and the philosophy of measurement. Four normative strands are presented to guide this rethinking: an emphasis on the logic of discovery alongside the customary logic of justification, a move towards radical transparency, an invitation to philosophical exploration, and an appeal to integrate qualitative analysis at every major step of the measurement process. The result of following these normative strands would be a discovery-based workflow for educational measurement. To illustrate such a workflow, I take as an example data from a program evaluation of an intervention for underrepresented minority students in a university STEM program.

Table of contents

List of Figures

List of Tables

Chapter 1

Introduction

"Neither the naked hand nor the understanding left to itself can effect much. It is by instruments and helps that the work is done, which are as much wanted for the understanding as for the hand. And as the instruments of the hand either give motion or guide it, so the instruments of the mind supply either suggestions for the understanding or cautions."

– Francis Bacon, Novum Organum, 1620

"Inchworm, inchworm
Measuring the marigolds, you and your arithmetic
You'll probably go far.
Inchworm, inchworm
Seems to me, you'd stop and see
How beautiful they are."

– Frank Loesser, children's song, 1952

1.1 Overview and Research Goals

Measurement occupies a central position in contemporary educational research. Statistical methods are currently the key means for testing generalizations about learning, teaching, funding, and organizational structure. These statistical methods, in turn, depend on thousands of measures of aptitudes, attitudes, and behavior developed and published by researchers each year. Measurement is responsible for no less than the interface between

reality and the statistical model. Decisions made at this level, while often concealed by opaque jargon and disciplinary divisions of labor, resonate throughout all subsequent applications of the model. What is not measured is no longer visible through the lens of the model, and what is measured poorly may be dramatically distorted. Good measures have the power to reveal social facts about the educational system, such as disparities in educational opportunities and the power of positive interventions. Poorly constructed measures leave us in ignorance or bolster falsehoods.

The theoretical issues posed by the measurement phase of educational research are among the most interesting in the field: How should we decide what to measure? What counts as a rigorous definition of that construct? How do we know that we have measured well? The field of academic research tasked with answering these questions across the sciences has been called the "philosophy of measurement" (Michell, 2005). Although young, the philosophy of measurement and other critical approaches have yielded invaluable insights into the domain of applied statistics known as psychometrics, which seeks to measure attitudes, aptitudes, and behaviors. These philosophical insights into psychometrics have predominately been framed in terms of validity, the definition of which has progressively expanded from the minimalistic extent to which “a test measures what it is supposed to measure” (Garrett, 1937, p.324; Loevinger, 1957) to a full framework for weighing logic and evidence pertaining to test content, response processes, internal mathematical structure, correlations with other measures, and the consequences of testing (AERA, APA, NCME, 2014).

It is now clear that the main challenges facing measurement are not issues of statistics but of research design. Statistical techniques have never been the only methods of assessing

validity, but they have always been the focus of validation. Methods such as item-paneling, think-alouds, cognitive interviews, card sorts, and concept mapping home in on the response process and push researchers to more precisely define the constructs under study (Wilson, 2004, Onwuegbuzie et al., 2010; Willis, 2015). Due to the renewed theoretical emphasis on validity, measurement is increasingly embracing a version of mixed methods research. The recommendation that the used of mixed methods would improve the development of measures has been commonplace in the last 15 years (Collins et al., 2006; Onwuegbuzie et al., 2010). However, Zhou (2019) has argued that in recent years "Although researchers argued the advantages of using mixed methods to develop new scales, literature providing systematic instructions on how to do it has been scarce and incomplete" (p.39). That is, *why* researchers should use mixed methods for educational measurement is now clear, even if *how* they should do so is not.

1.2 The Case for a Workflow

The practical and theoretical challenges facing educational measurement today seem unlikely to be addressed by the addition of new rules and guidelines to the Standards: even the existing rules are regularly disregarded by researchers, the state, and testing corporations. Rather, what would seem to be most helpful for practitioners would be a theoretically and practically informed example workflow that could be adapted to serve the needs of many projects. This workflow would, ideally, prompt practitioners to address major theoretical issues in educational measurement, combine the methodological expertise embodied in both qualitative and quantitative traditions, and embody best practices of analysis where available. The decision to contribute a "workflow" rather than a set of prescriptions for practitioners or

a formalization of a putative underlying structure of educational measurement is a considered one. Workflows are progressions of processes in which each phase builds logically upon the previous phase. They are a staple of the creative process rather than a form of judgment or critique, although enhancing the quality of work is one of their essential functions. A good workflow transforms raw effort, talent, and knowledge into a viable product without wasting these resources along the way (Meir, 2018).

The workflow in this dissertation is suggested as a guide to future travelers along the path of educational measurement. Guides are only meant to improve the journey. To begin, I consider several existing workflows for the construction of educational measures. Then, I synthesize the best features of these workflows into my own workflow, addressing their shortcomings along the way. The remainder of the dissertation consists in a worked example of the development of a measure using this workflow. The example case is of a self-report survey about academic habits, and the specific steps of the workflow are certainly adequate for the construction of similar instruments. However, while the general phases and principles of the workflow easily apply to the construction of different kinds of educational measures, such as classroom assessments, some of the specific steps may not. Every measurement situation is unique in some ways, and it up to the reader to determine how best to adapt this workflow for their own needs.

Mapping out this workflow is not an attempt to add another methodological hurdle over which practitioners must vault in order for their research to be considered rigorous, trustworthy, or scientific. Cronbach and colleagues (1980) were undoubtedly correct when they said that "Merit lies not in form of inquiry, but in relevance of information" (p.7). Instead, this workflow embodies one specific answer to several enormous questions in

measurement. For those who are exercised by these enormous questions, it is meant to provide a workable alternative to common workflows for the creation of educational measures. A good measurement workflow should produce a great quantity of "relevant information", and this information should be of such a quality that new discoveries are possible.

1.3 Common Workflows for Educational Measurement

1.3.1 Evidence Centered Design

Perhaps the most widely known workflow for the construction of educational measures is that of evidence-centered design (Mislevy et al. 2003). Evidence-centered design (ECD) was developed by researchers at the Educational Testing Service and is thus used primarily for the construction of assessments, although there is nothing in principle to prevent it from being used for the development of behavioral and psychological measures. The primary focus of the ECD workflow is the use of evidentiary reasoning (Mislevy, Almond, & Lukas, 2003) - that is, the development of measures whose primary characteristics can be defended by appeal to an evidence base, drawing heavily on the theoretical work of Messick (1989, 1994) and Kane (1992). The latter, in particular, proposes that the inferences underlying the development of any measure can be conceptualized as an elaborate logical argument, an approach known as the "argument-based approach" to validity (Kane, 1992). Mislevy and colleagues (2002) envision a similar logical scaffold supporting the use of instruments, arguing that:

> [F]lexible models and powerful statistical methods alone are not good enough. It is a poor strategy to hope to figure out “how to score it” only after an assessment has been constructed and performances have been captured. Rather, one should design a

> complex assessment from the very start around the inferences one wants to make, the observations one needs to ground them, the situations that will evoke those observations, and the chain of reasoning that connects them (p.364).

The workflow of Evidence Centered Design consists four large phases: Domain Analysis, Domain Modeling, the Conceptual Assessment Framework, and the Operational Assessment. Domain analysis consists of "marshaling substantive information about the domain—bringing together knowledge from any number of sources and then beginning to organize beliefs, theories, research, subject-matter expertise, instructional materials, exemplars from other assessments, and so on" (Mislevy, Steinberg, & Almond, 2003, p.7). Domain Modeling is the organization of this information into paradigms and structural relationships. The Conceptual Assessment Framework is the blueprint for the assessment. It consists of three conceptual models: the Student Model, the Evidence Model, and the Task Model. The Student Model is the set of relationships between knowledge, skills, and abilities that are relevant to the measurement of some outcome, such as "algebra proficiency." The Evidence Model connects observations of human behavior to the variables identified in the Student Model, and includes rules for handling evidence and the statistical model. The Task Model consists of specifications for the part of the measure with which the student will interact directly, such as characteristics of the stimulus and instructions. The final developmental stage, the Operational Assessment, is concerned with in the presentation, task selection, and scoring of the measure.

Evidence centered design is the most comprehensive workflow for educational measurement and has much to recommend it. ECD emphasizes the importance of logical steps, slowing down a process that is typically undertaken with too much haste. These steps leave behind a transparent record of major decisions that can be consulted later to clarify the

intended relation between intentions and design. By separating the process into so many discrete stages, the progressive logic of moving from one stage to the next can be more easily checked, and the strength of individual parts of the model can be evaluated.

1.3.2 The BEAR Assessment System

The BEAR[1] Assessment System (BAS) is another workflow for the development of educational measures (Wilson & Sloane, 2000). Like Evidence Centered Design, the BAS was conceived primarily with assessments rather than psychological scales in mind - however, unlike ECD, BAS has been rigorously applied by its creators to psychological (Dawson et al., 2010; Liu, & Wilson, 2010), sociological (Rocca, Krishnan, Barrett, & Wilson, 2010), and medical (Wilson, Allen & Li, 2006) constructs. The BAS is anchored firmly within the framework of Rasch Measurement Theory and draws theoretical strength from the inclusion of Rasch-specific data representations and model semantics. The BAS was codified as a workflow for the development of measures in Wilson's *Constructing Measures* (2004). This workflow consists of four building blocks: construct definition, item design, outcome space, and measurement model. Construct definitions are formalized as construct maps, orderings of qualitatively different levels of performance that may be evinced by participant choices and behavior. Items design refers to the process of creating items or tasks that elicit performances corresponding to levels on the construct map. The outcome space is the set of categories into which performances are organized, such as low to high levels of mastery of a domain of knowledge or bronze to gold Olympic performances. The measurement model refers to the statistical model used to estimate participant's level on the

[1] BEAR stands for the Berkeley Evaluation and Assessment Research (Center) - making BAS an acronym containing another acronym.

construct. The model is formalized in statistics about item characteristics (e.g. item fit, differential item functioning) and the data representation known as a Wright Map, which directly compares estimates of participants' level on the construct to thresholds of severity or difficulty embodied by the items.

Where ECD is maximalist, the BAS is minimalist. The above description includes the only four steps in the workflow, which is meant to be pursued iteratively until the desired level model fit, reliability evidence, and validity evidence has been achieved (Wilson, 2004, p.19). This minimalism makes the BAS easier to teach and learn than ECD. The BAS also contains a more prescriptive set of recommendations than ECD, such as normative arguments for cognitive interviews, item panels, finite and ordered categories in the outcome space, and the use of Rasch measurement theory (Wilson, 2004). While less flexible, the BAS thus is an easier workflow to follow from a practical standpoint.

1.3.3 Sequential Mixed Methods Designs

To the aforementioned workflows for the construction of measures in education, we might add several from outside the field as well. In their landmark guide to mixed methods designs, Creswell and Plano Clark (2011) outline the following scenario as ideal for mixed methods research: "A researcher seeks to evaluate a program that has been implemented in the community. The first step is to collect qualitative data in a needs assessment to determine what questions should be addressed. This is followed by the design of an instrument to measure the impact of the program." In their instructions for carrying out this task, the authors adapt the phases of scale development from Devillis (2012), who lays out the following steps:

> 1. Determine what you want to measure and ground yourself in theory and in the constructs to be addressed (as identified by the qualitative findings). 2. Generate an item pool, using short items, an appropriate reading level, and items that ask a single question (based on participant language identified in the qualitative findings when possible). 3. Determine the scale of measurement for the items and the physical construction of the instrument. 4. Have the item pool reviewed by experts (such as participants from the qualitative phase who are experts in their own experiences in addition to formally trained experts). 5. Consider the inclusion of validated items from other scales or instruments to detect undesirable responses. 6. Administer the instrument to a development sample for validation. 7. Evaluate the items (e.g., reverse scoring, item-scale correlations, item variances, factor analysis, coefficient alpha reliability, analysis of participant comments). 8. Optimize scale length based on item performance and reliability checks (Creswell & Clark, 2011).

These instructions have several virtues: they alert the researcher to decisions that will have to be made, refer to more than one way to use qualitative data, and urge recurring contacts with members of the focal population. Creswell and Plano Clark have already made major improvements over the steps given by Devellis (2012) in the latter's book on scale development, which includes roughly two pages on the use of qualitative methods.[2] The authors also point to several studies which offer more detailed accounts of the integration of qualitative methods into the measurement process (e.g. Betancourt et al., 2011; Meijer, Verloop, & Beijaard, 2001; Cinamon & Dan, 2010; Sinley & Albrecht, 2016). A commonality among these studies is their careful treatment of the item authorship process. For example, Meijer and colleagues (2001) performed and published a qualitative study of the construct, which was used to generate a taxonomy of the phenomenon (teachers' practical knowledge about reading comprehension). Items were each formulated using participant language and in consultation with participants through multiple phases of the instrument

[2] In addition, Creswell and Clark's listing of CTT methods for measurement is likely informed by Devellis, who argues against the replacement of CTT by IRT methods in the Fourth Edition of his *Scale Development* (2012). The argument hinges on the premise that IRT is not "necessarily superior" to CTT (p.216), while acknowledging that IRT has clear theoretical and practical advantages.

development process (Meijer, Verloop, & Beijaard, 2001). A second commonality is a more-than-typical concern for the validity of the measures involved.[3] Betancourt and colleagues (2011), for example, evaluated criterion validity by comparing the results of their measures to the diagnoses of local Rwandan health officials. Such iterative qual-quant-qual designs take maximum advantage of what Newman and Benz (1998) have called the “interactive continuum” of mixed methods research.

1.3.4 The Ethnographer's Toolkit

Researchers who identify primarily as ethnographers have also suggested mixed-methods workflows for the development of instruments. Schensul and LeCompte's (2012) thorough seven-volume *Ethnographer's Toolkit* dedicates a chapter to the topic. The authors place a heavy emphasis on the use of ethnographic data as a source for constructs and items. Ethnographic data is used to identify domains, each with a subset of "factors" (used here in the non-statistical sense), which are further subset into variables. The recommended process for accomplishing this task involves the construction of what amounts to a wall-sized map of factors, which are then used as headers for variables of interest related to that factor (e.g Domain - work; Factor - job satisfaction; Variable: adequacy of salary). As the process of instrumentation is carried out, some of the factors are eliminated and some form the basis for scales that will be piloted with the target population. This strategy of construct selection is clearly meant to proceed from a very wide scope to a much narrower one. The task is to reduce the number of potential areas of consideration, rather than "inventing" them (Book 3,

[3] Validity remains a central concern in these studies even when the technical language of validity is not invoked. In contemporary terminology, these researchers expended the most effort on gathering evidence of the validity of content.

2013, p.256). Once the instrument is ready to be piloted, fairly typical methods of statistical analysis take over in the construction of the scale.

There is obviously much of value in Schensul and LeCompte's workflow. The emphasis on beginning with ethnographic data and constantly referring to it or gathering additional qualitative data through the process of measure development is unique and principled. The authors pinpoint the importation of generic questionnaires that are "validated" using other populations as a highly problematic practice from an ethnographic point of view, and insist on the necessity of local adaption and validation. Several provocative, bright lines are drawn:

> Ethnographic surveys are *never*: the first data collection operation in a research project; the product of a process in which research staff generate survey items based only on their own personal experience or only on existing instruments...; opportunities to limit a study to the use of standardized indices from other studies and/or validated on national samples...; part of the 'discovery' process. (Book 3, 2013, p.244).

Further, the items selected as part of an instrument should be derived from ethnographic data that reflect the "identification and linking of domains/factor/variable hierarchies or taxonomies relevant to the local setting" (Book 3, 2013, p.244). That is, all instruments used should only measure constructs identified by qualitative research.

1.4 Limitations of Existing Workflows

Workflows are more than mere practical tools: like all forms of *techne*, workflows embody implicit and explicit *epistemes*. All workflows for the development of measures imply a process of knowledge construction. Unlike many tools of educational research, once created, measurement instruments may remain unchanged for decades, carrying forward any limitations they may possess. While each of these workflows for the development of

educational measures is very helpful in its own way, there remains something to be learned by turning a critical eye towards each of them.

I suggest that the primary weaknesses of the ECD workflow lie in the early processes of selecting constructs for measurement. This weakness is notable in two elements of the ECD process - domain analysis, which is the first phase in ECD, and the creation of the Student Model. Domain analysis includes a set of guidelines for drawing out the features the construct, such as an examination of knowledge valued in real-world situations (Mislevy, Steinberg, & Almond, 2003, p.18). However, it is easy to miss the fact that these guidelines do not refer to the selection of which constructs to measure, but rather to gaining better understanding of the constructs that have already been selected. Indeed, in the published examples of ECD workflows, construct selection always proceeds primarily via consultation with disciplinary experts (Mislevy et al., 2002; Mislevy, Steinberg, & Almond, 2003; Mislevy, Almond, & Lukas, 2003). There is a thoroughgoing pluralism at the heart of ECD, which is content to begin the measurement process from virtually any standpoint:

> A conception of the knowledge or skill one wants to measure can be a useful starting point for assessment design... So can the kinds of things one wants to see students learn to do... or real-world situations in which we are ultimately interested... These points of view correspond to an emphasis on concerns that are central to the proficiency, evidence, and task paradigms. Good assessment comes not from "choosing the right one" but by synthesizing these concerns. Creating a collection of interlocking paradigms ensures that the elements which are highlighted in the different perspectives have been thought through and integrated (Mislevy, Steinberg, & Almond, 2003, p.8).

While it is laudable that reasons for construct selection beyond the goals of the assessor are mentioned in discussions of ECD, I argue that such pluralism is extremely permissive as a foundation for measurement and clouds the relation between observation and theory in

domain analysis. In ECD, it appears that nearly any grounds may serve as the justification for the measurement of any construct. In a separate introduction to ECD, Mislevy and colleagues (2002) take a position of deliberate openness regarding the grounds of the Student Model:

> "Configurations of values of student model variables (SM variables) are meant to approximate certain aspects of the infinite configurations of skill and knowledge real students have, as seen from some perspective about skill and knowledge in the domain. It could be the perspective of behaviorist, trait, cognitive, or situative psychology. This perspective determines the kinds of stories we want to weave for our purposes" (p.367).

This emphasis on the "stories we want to weave" occurs throughout the article. Arguably, this approach is a formalization, in the construct selection phase, of the narrative fallacy: choosing a compact story that explains events and overinterpreting the causal factors that allegedly make a difference in this story (Taleb, 2007, p.63). I discuss the ECD criteria for construct selection in greater detail in the subsequent chapter. At present, it suffices to explain that openness to virtually any criteria for construct selection is a *de facto* abdication of criterial reasoning. For example, tradition ("we've always done it this way") or bureaucratic exigency ("it would be easier for the administration") may be sufficient justifications for selecting constructs for educational measurement according to ECD - if they are insufficient, it is unclear why this would be the case.

The primary limitation of the BEAR Assessment System rhymes with the above discussion of ECD. The first step in the BAS is the construct definition phase, which begins with the drawing of an item map. The BAS is agnostic to the origin of the notion of the particular construct, placing the workflow in the rationalist, Cartesian tradition. To borrow a Newtonian phrase, such rationalist approaches risk "feigning hypotheses" - that is, postulating entities or processes for which we lack independent evidence (Janiak, 2010).

Moving so quickly to construct definition begins the process of measurement construction with a reification of the construct. Arguably, the later checks in the BAS workflow, such as cognitive interviews and item paneling, might uncover falsificatory evidence, but these steps as portrayed in *Constructing Measures* are more focused on design improvements to the instrument. Thus, whereas ECD is open to construct selection using virtually any process, BAS assumes that the construct has already been selected or that it can be rationally deduced.

The limitations of the Cresswell-Clark-Devellis workflow are both conceptual and statistical. First, some instructions signal decisions to be made rather than any means or criteria of making them. This is analogous to a recipe which prescribes "bake the cake" without mention of the necessary oven temperature or baking time. From Steps 1-3, we learn about the kinds of decisions that need to be made (Construct Selection, Instrumentation), but the relation between inputs (qualitative data) and outputs (decisions) remains unclear. Second, these instructions commit implicitly to atheoretical approaches to measurement, such as the deletion of items based on statistical rationales alone. Third, these instructions remain within the measurement paradigm of classical test theory, which has been theoretically and mathematically eclipsed by item response theory since the 1980's (Jaeger, 1987; Embretson, 1996). Perhaps most troubling from a disciplinary perspective is the fact that these instructions, taken together, may be construed as an endorsement of the lowest standards for measure construction in use today. And yet, these limitations are in no way confined to Creswell and Clark, who represent some of the leading voices urging the integration of qualitative methods into measurement. Each of these example studies mentioned by Cresswell and Plano Clark represent at least one of the above shortcomings implied by their adaptation of Devellis' scale development process. All employ some version of atheoretical

scale shortening, such as deleting items that do not load onto principal components in a PCA or to increase Cronbach's alpha, an approach sometimes undertaken without further regard to the dimensionality of the scale. None employ item-response theory models or employ model-based reasoning about basic measurement issues such as construct representation (Messick, 1995). The fact that these exemplary studies feature a wide range of statistical sophistication suggests that these shortcomings result from limitations of theoretical perspectives rather than a lack of technical knowledge. This theoretical limitation might be summed up as an incomplete grasp of the theoretical challenges of human measurement. In methodological terms, this limitation leads to an as-yet incomplete integration of mixed methods processes at the measurement phase.

As for Cresswell and Clark above (2011), it is instructive to consider some of the conspicuous absences in Schensul and LeCompte's workflow that may be pertinent from a measurement perspective. As Book 5 of the *Toolkit* makes clear, the ethnographic surveys created using the workflow in Book 3 will eventually be used as indicators of some target construct, including in sophisticated statistical designs. Construct selection in the *Toolkit* is discussed primarily as a practical process rather than as a theory-laden decision that requires its own warrants. We might call the suggested approach a "construct meta-map" since it involves the simultaneous mapping of multiple potential constructs observed in the field. As a knowledge representation, this is edifying. However, the rationales for pruning of constructs from the meta-map come out to an odd mixture, including 1) a rule of thumb to select between two and five independent domains and one or two dependent domains, 2) the deletion of "factors" with relatively fewer "variables" in the meta-map, 3) the deletion of "variables" from relatively crowded "factors", 4) the deletion of "domains" with relatively

fewer "factors", 5) deferring to the judgment of the project leader to determine deletion. Unfortunately, the arbitrariness of several of these practices from a measurement perspective undercuts the utility of the meta-map. Several of these decisions appear more aesthetic than theoretical, subtly shaping "goldilocks" meta-maps featuring a balanced appearance and a medium number of factors for each subset. A researcher following these guidelines is quite liable, for example, to prune away the most important predictors of a substantive outcome, such as academic performance or drug recovery, if these do not split analytically into a pleasing number of factors or variables. The second major limitation from a measurement perspective is the absence of any discussion of how measurement principles can inform the construction of an instrument. Stevens' levels of measurement are introduced (nominal, ordinal, interval, and ratio), but only nominal and ordinal levels are treated at length, with the usual caveats that "the distances between levels do not have any meaning" in ordinal measurement (p.261).[4] The authors come frustratingly close to handling this problem - which item response theory is designed to address - when they introduce Guttman scales and their potential for ranking participants, even parenthetically citing an introduction to the Rasch model by Andrich (1985). Cronbach's alpha and factor analysis are briefly mentioned in Book 5 of the *Toolkit* (pp.197, 202), although the introduction of these tools emphasizes the development of conventionally adequate scales rather than the logic of measurement.

[4] The authors of the *Toolkit* appear oddly hung up about the existence of real or assumed zero points in interval-level measures, implying that this is a major problem in Book 2 and Book 5, e.g. "It is difficult to find true interval variables in social science research that are not based on a real or assumed zero start point." The presence or absence of a zero point is not considered a serious limitation to the approximation of interval-level measurement in item-response theory.

1.5 Proposed Workflow

In this dissertation I propose four normative strands of improvement to these workflows for the development of educational measures. First, I suggest the adoption of an epistemic posture oriented towards discovery. Kuhn has famously argued that the historical distinction between the "context of justification" and the "context of discovery" is a false dichotomy (Kuhn, 1996, p.8). Some measurement workflows, such as those that employ the argument-based approach, work within the context of justification (Kane, 1992). Others might be built that attend equally or preferentially to the context of discovery. This difference can be conceived of as a continuum: the most discovery-based workflows would take in a broad variety of information and avoid funneling it too quickly through an analytic process, while the most justification-based workflows would attempt to build the strongest model first and then accumulate evidence that nothing untoward had occurred in the process - an intermediate process might seek disconfirming evidence at any number of stages. The idea of a discovery-based workflow is prefigured in some of the authors already mentioned, although its implications for measurement have perhaps yet to be fully explored. Wilson (2004) speaks of "the degrees of prespecification" of item formats, with the most open-ended methods - he includes observation and interviews - logically being used at the beginning of the research project and fixed response items being employed only when the outcome space has been suitably defined (pp.50-52). The *Ethnographer's Toolkit* likewise mentions a similar progression along a "ladder of abstraction" from observation and unstructured interviews to methods that elicit and curtail participant behavior more markedly, such as structured interviews and surveys (Book 3, p.248). A discovery-based workflow would move from

methods that are maximally open-ended to methods that tightly define constructs, instruments, and claims.

The second normative strand I suggest is radical transparency. Too many educational measures lack sufficient evidence to be fully evaluated or adapted for later use. As the Chinese Legalist statesman Han Feizi noticed, kings are more powerful when they cultivate an air of mystery about their true character and desires (Watson, 2003). Paradoxically, hiding the details of measure development can invest the process with greater power and legitimacy than warranted.

The third normative strand in this workflow is the deliberate integration of philosophical exploration of the construct into the workflow. Developing measures is a philosophically-demanding activity and it should be treated as such (Alexandrova, 2017). Critical epistemic and ethical questions should be asked about the decisions we make at each step of measure construction.

The fourth normative strand in this workflow, meant to amplify the potential of the previous three strands, is the injunction to integrate qualitative methods at every phase of measure development. In a review of literature from the 2010's, Zhou (2019) found that the "qualitative methods" employed in early phases of measure development were most often consultations with experts and literature reviews, rather than the signature methodologies of qualitative research, such as ethnographic observation or semistructured interviews. In education, many mixed methods approaches integrate qualitative methods at the validation stage (Burton & Mazerolle, 2011; Nassar-McMillan, Wyer, Oliver-Hoyo, & Ryder-Burge, 2010; Smolleck et al., 2006). While mixed methods are of course important for validation, I argue that qualitative data has a major role to play in all phases of instrument development.

This argument echoes a larger discussion within the mixed methods literature about the limitations of strictly sequential designs, which do not benefit optimally from the complementarity of mixed methods (Newman et al., 1998; Leech & Onwuegbuzie, 2010; Onwuegbuzie et al., 2010; Castro, Kellison, Boyd, & Kopak, 2010).

In order to instantiate these four normative strands in the workflow of measure development, I suggest a four-step workflow for measure development: construct selection, construct definition, instrumentation, and validation. In the construct selection phase, descriptive evidence should be provided that the construct meets a defensible constellation of criteria for measurement. In the construct definition phase, the conceptual scope of the construct is fully explored and a normative argument for measuring the construct is stated. In the instrumentation phase, insights from the previous two phases are carried forward in the construction of an instrument. In the validation phase, the instrument is subjected to evaluation of its quality and improvements are suggested. I argue that these four phases bring together the best elements of the previously mentioned workflows for the development of educational measures. For educationalists seeking to answer the need to "measure something" this workflow embodies many evidential and normative commitments, offering checkpoints to avoid historical mistakes of educational measurement.

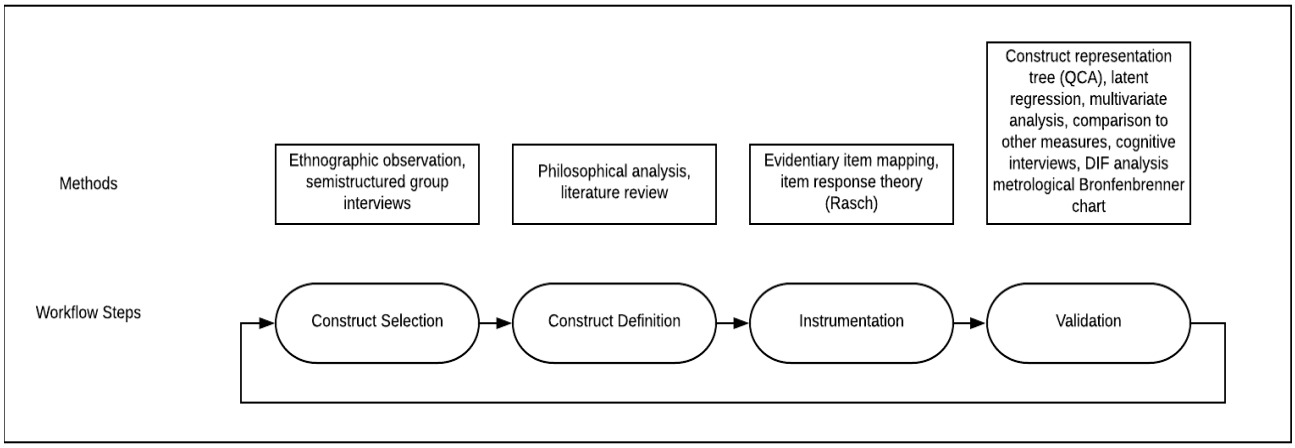


Figure 1.1: Steps in the discovery-based workflow and methods chosen in this dissertation to correspond to each step. The workflow may cycle back to the beginning if necessary so that the processes in each step can be revisited.

These four logical steps do not correspond to familiar measurement paradigms such as the BEAR assessment system (Wilson, 2004) or evidence-centered design (Mislevy et al. 2003), but they are indebted to them. In my conceptualization, the "construct mapping" phase in the BEAR system and the "domain analysis" phase in ECD are deliberately split into the construct selection and construct definition phases in order to emphasize the importance of the selection phase and the importance of arriving at a definition. This seems the only way to satisfy the normative philosophical strand of improvement above. The rest of the steps of the BEAR system (item design, outcome space, and measurement modeling) and the two following steps of ECD (domain modeling, conceptual assessment framework) are summarized in the instrumentation phase. The process of measure construction as I understand it concludes with instrument validation as the final phase, in order to emphasize that a measure has not actually been fully developed until the validation process has come to

a satisfactory stopping point. The *differences* between my conceptualization and alternative workflows are not the focus of this dissertation, and it is probable that users of those frameworks would agree that all four of these stages are logically necessary or that these stages are consistent with their approach. Rather, the focus of the dissertation is synthesis of the best available techniques into an approachable and intelligible workflow for researchers. Pieces of the framework I propose in this dissertation are scattered throughout the literatures I have consulted in education, measurement, survey design, and anthropology.

Each of the next four chapters are each dedicated to carrying through one of the four steps using a case study from a program evaluation. Among the various types of mixed methods designs, the present study is an instance of the "survey-development variant" of an exploratory sequential design (Creswell & Plano Clark, 2011). This is to say that it is a member of the larger class of exploratory sequential designs - a popular research design in which early phases of research are used to narrow down research questions that will be the focus of later phases. A special case of the exploratory sequential design involves the development of an instrument, conceived as a middle phase. Instruments that are designed using qualitative inquiry "offer a key connection between the primary methodologies" (Creswell & Plano Clark, 2011). This connective potentiality is a major motivation for the normative strand of including qualitative methods at every step. By beginning the workflow on a strong qualitative foundation, the workflow provides for a nutritive basis of information - the albumen of the egg - on which the research can draw to develop the measure.

References

Alexandrova, A. (2017). *A Philosophy for the Science of Well-being*. Oxford University Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*.

Andrich, D. 1985. An elaboration of Guttman scaling with Rasch models for measurement. In *Sociological methodology*, ed. N. B. Tuma, 33–80. San Francisco, CA: Jossey-Bass.

Betancourt, T. S., Meyers-Ohki, S. E., Stevenson, A., Ingabire, C., Kanyanganzi, F., Munyana, M., . . . Beardslee, W. R. (2011). Using mixed methods research to adapt and evaluate a family strengthening intervention in Rwanda. *African Journal of Traumatic Stress*, 2(1), 32–45.

Burton, L., & Mazerolle, S. (2011). Survey instrument validity part I: Principles of survey instrument development and validation in athletic training education research. *Athletic Training Education Journal*, 6(1), 27–35.

Castro, F. G., Kellison, J. G., Boyd, S. J., & Kopak, A. (2010). A methodology for conducting integrative mixed methods research and data analyses. *Journal of mixed methods research*, *4*(4), 342-360.

Cinamon, R. G., & Dan, O. (2010). Parental attitudes toward preschoolers' career education: A mixed-method study. *Journal of Career Development*, 37(2), 519–540.

Collins, K., Onwuegbuzie, A., & Sutton, I. (2006). A model incorporating the rationale and purpose for conducting mixed-methods research in special education and beyond. *Learning Disabilities: A Contemporary Journal*, 4(1), 67–100.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.

Dawson-Tunik, T. L., Goodheart, E. A., Draney, K., Wilson, M., & Commons, M. L. (2010). Concrete, Abstract, Formal, and Systematic Operations as Observed in a "Piagetian" Balance-Beam Task Series. Journal of Applied Measurement, 11, 1.

DeVellis, R. F. (2012). *Scale development: Theory and application* (3rd ed.). Newbury Park, CA: Sage.

DeVellis, R. F. (2017). *Scale development: Theory and application* (4th ed.). Newbury Park, CA: Sage.

Embretson, S. E. (1996). The new rules of measurement. *Psychological assessment*, 8(4), 341.

Garrett, H. E. (1937). *Statistics in psychology and education*. New York: Longmans, Green.

Jaeger, R. M. (1987). Two decades of revolution in educational measurement!?. *Educational Measurement: Issues and Practice*.

Janiak, A. (2010). *Newton as Philosopher*. Cambridge University Press. Cambridge, England.

Kane, M.T. (1992) An argument-based approach to validity. Psychological Bulletin, 112, 527-535.

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.). University of Chicago press.

Leech, N. L., & Onwuegbuzie, A. J. (2010). Guidelines for conducting and reporting mixed

research in the field of counseling and beyond. *Journal of Counseling & Development*, 88(1), 61-69.
Liu, O. L., & Wilson, M. (2010). Sources of Self-Efficacy Belief: Development and Validation of Two Scales. *Journal of Applied Measurement*, 11(1), 24-37.
Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological reports*, 3(3), 635-694.
Meijer, P. C., Verloop, N., & Beijaard, D. (2001). Similarities and differences in teachers' practical knowledge about teaching reading comprehension. Journal of Educational Research, 94(3), 171–184.
Meir, D. (2018). *Workflow: A Practical Guide to the Creative Process*. CRC Press.
Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.
Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9), 741.
Michell, J. (2005). The logic of measurement: A realist overview. *Measurement*, 38(4), 285-294.
Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003a). A brief introduction to evidence-centered design. ETS Research Report Series, 2003(1), i-29.
Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003b). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-67.
Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (2002). Making sense of data from complex assessment. *Applied Measurement in Education*, 15, 363-378.
Nassar-McMillan, S., Wyer, M., Oliver-Hoyo, M., & Ryder-Burge, A. (2010). Using focus groups in preliminary instrument development: Expected and unexpected lessons learned. *The Qualitative Report*, 15(6), 1621–1634.
Newman, I., Benz, C. R., & Ridenour, C. S. (1998). *Qualitative-quantitative research methodology: Exploring the interactive continuum*. SIU Press.
Onwuegbuzie, A. J., Bustamante, R. M., & Nelson, J. A. (2010). Mixed research as a tool for developing quantitative instruments. *Journal of Mixed Methods Research*, 4(1), 56-78.
Rocca, C. H., Krishnan, S., Barrett, G., & Wilson, M. (2010). Measuring pregnancy planning: An assessment of the London Measure of Unplanned Pregnancy among urban, south Indian women. *Demographic research*, 23, 293.
Schensul, J. J., & LeCompte, M. D. (2012). *Ethnographer's toolkit.* Rowman Altamira. Walnut Creek, CA.
Sinley, R. C., & Albrecht, J. A. (2016). Understanding fruit and vegetable intake of Native American children: A mixed methods study. *Appetite*, 101, 62–70.
Smolleck, L., Zembal-Saul, C., & Yoder, E. (2006). The development and validation of an instrument to measure preservice teachers' self-efficacy in regard to the teaching of

science as inquiry. *Journal of Science Teacher Education*, 17, 137–163.
Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. Random house.
Watson, B. (2003). *Han Feizi: basic writings*. Columbia University Press.
Willis, G. B. (2015). *Analysis of the cognitive interview in questionnaire design*. Oxford University Press.
Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Routledge.
Wilson, M., Allen, D. D., & Li, J. C. (2006). Improving measurement in health education and health behavior research using item response modeling: comparison with the classical test theory approach. *Health education research*, 21(suppl_1), i19-i32.
Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied measurement in education*, 13(2), 181-208.
Zhou, Y. (2019). A Mixed Methods Model of Scale Development and Validation Analysis. *Measurement: Interdisciplinary Research and Perspectives*, 17(1), 38-47.

Chapter 2

Construct Selection

2.1 Introduction

In educational measurement, it is frequently the case that 1) we desire to measure some subset of attributes of learners, teachers, or learning situations, and 2) there is no definitive boundary or list of what these attributes might be. Often, a further condition obtains, namely that 3) measuring at least one of the potential attributes in the subset would require the development of a novel instrument. This chapter is concerned with addressing conditions 1 and 2, while subsequent chapters are focused on addressing condition 3. I argue that the situation implied by the conjunction of the first two conditions is so common as to be a near constant in educational research, yet this issue has rarely been seriously treated as a measurement issue. As I will show, a sense of overconfidence about the ability to resolve these two conditions may arise from certain misconceptions about the way that familiar tools actually work. Ultimately, the issue of how to address these first two conditions is reframed as an *empirical* problem: how can construct[5] selection proceed using empirical evidence for the decision rather than non-empirical forms of reasoning. Ethnographic data is used by some educational researchers to develop measures (Crede & Borrego, 2013; Hitchcock et al., 2006; Nastasi et al., 2007). I suggest some criteria for empirical construct selection from ethnographic data. The process of reasoning about construct selection is illustrated using a case study from a goal-free (Scriven, 1973) program evaluation of an intervention for first-

[5] A brief note on language. "Attribute" is used to denote the wide variety of real things one might attempt to measure. "Construct" is used to denote the hybrid objects (Latour, 2012) that bear the obvious imprint of research methodology, qualitative or quantitative. As Jane Loevinger quipped in 1957, "Traits [or attributes] exist in people; constructs (here usually about traits) exist in the minds and magazines of psychologists." Following Loevinger, the difference between attributes and constructs is analogous to the mathematical distinction difference between parameters and statistics.

year biology students at a large public university.[6] I begin with an ethnographic description of the talk and behavior of students in this program, and then proceed to the discussion of the methodological problem of selecting a construct for measurement. This process of the Construct Selection is the first phase of a discovery-based measurement workflow: that is, a workflow for empirical work that seeks to discover appropriate constructs to measure in real-world educational situations.

2.2 Program

The case study analyzed in this dissertation is part of an evaluation of a successful academic program for freshmen known as MCDB-11 or BIOME (Biology Mentoring and Engagement).[7] The goal of this program is to recruit under-represented minorities (URMs)[8] in the pre-biology major and provide support that will contribute to their retention. Retention of URMs in STEM majors is a local as well as national STEM pipeline issue (National Center for Education Statistics, 2005; DePass and Chubin, 2009; Estrada et al., 2016). Internal program documents describe the curriculum as encompassing "a variety of topics, including time management, looking at science critically, and exploring various career options," with the intent "to help freshmen transition into successful college students." The

[6] Goal-free evaluation program evaluation is characterized by a rejection of the traditional paradigm of program evaluation, which is to define program "goals" in advance and then select appropriate measures of these goals. The case against simply following stated goals typically involves an acknowledgement that the concept of an evaluation goal is fraught with its own complications. As Cronbach and colleagues stated in 1980 "Goals are a necessary part of political rhetoric, but all social programs, even supposedly targeted ones, have broad aims." Not all goals are stated, and not all the stated goals will actually be pursued. Observing the operations of the program in order to formulate a mechanism of its functioning is one alternative to stating goals in advance.

[7] The present project was conducted with the permission and support of Dr. Mike Wilton, Department of Molecular, Cellular, and Developmental Biology at UC Santa Barbara. I am grateful for extensive access granted to me by Dr. Wilton, without which this project would not have been possible. Likewise, my collaborator in the program evaluation has been Dan Katz, whose insights have been invaluable.

[8] In STEM educational contexts, "underrepresented minorities" conventionally includes the African-American, Latinx, Native American, and Alaskan Native students.

110 first-year students enrolled in BIOME meet once per week for one hour with an upperclassman mentor and an instructor, an approach modelled on previously published mentorship programs (Otero et al., 2010; Solanki et al., 2019; Xu et al., 2018). Each session of BIOME is composed of six mentors with approximately five freshmen assigned to each, in a class of approximately 30 students. Students complete readings and small assignments in between sessions. Whole-class lessons are led by a biology professor, while table discussions are led by mentors.

In a prior cross-sectional study, students enrolled in this class showed a statistically significant gain in GPA of half a point on the 4-point scale when compared to a matched group of students who did not participate (Wilton et al., 2019). However, a causal mechanism (or set of mechanisms) that might account for the success of students in the program had not yet been identified. Classroom ethnography was chosen as the first step to explore this causal mechanism (Maxwell, 2004) *in situ*, with the understanding that this process would culminate in the construction of a survey measure to explore the generalizability of our hypotheses using a comparison group design with a larger sample of first-year students at the university - a sequential mixed methods design (Creswell & Clark, 2011) using an ethnographic survey (Schensul & LeCompte, 2012). In program evaluation terms, this was to be a goal-free evaluation (Scriven, 1973), meaning that explicit targets for evaluation had not been handed down from the program head, as well as a process-improvement evaluation (Chen, 1996), meaning that the goal of the evaluation was learning about and improving the program rather than making a summative judgment about its overall effectiveness. Thus, measurement was always in the goal, although it was unknown what the operative constructs might be and which of them we should select.

2.3 Qualitative Methods

2.3.1 Method 1: Classroom Observation

To understand how BIOME worked, I chose to attend the same 10-week series of classroom meetings that a first-year undergraduate would attend. I sat in the back of the room to one side, with my back to the wall. The structure of the lessons in BIOME alternated between whole-class discussion, which I could always hear and see, and discussions at individual tables, for which my observations were partly obstructed due to classroom noise. To observe table talk, I selected a table about three feet in front of me in the back right-hand corner and observed their conversations. Counting clockwise from the front of the room, I named this group "Table 3" and assigned a pseudonym to each participant, which I use throughout my field notes. The selection of a single group to closely observe over ten weeks offered the possibility of more detailed knowledge of intraindividual trajectories through the course, knowledge which would have been unattainable had I chosen to observe a different table each day. Students at Table 3 never varied their seating arrangement in the 10-weeks of observation except in the case of absences, so the diagram of the seating arrangement in Figure 5 applies to every day of observation, offering perhaps an unusual degree of consistency for an ethnographic setting. While the participants in the classroom were certainly aware of my presence, I asked the instructor to refrain from drawing attention to me, and at no point did my presence become conspicuous. During the ten class sessions I observed, no student or teacher spoke to me, and to my knowledge I was seldom looked at. To remain as unobtrusive as possible, I did not use field recording devices or a computer in the classroom, instead relying on a steady stream of jottings in my notebook, assisted by a

digital wristwatch to record the time at which key events occurred. In accordance with field note recommendations offered by Emerson et al. (2011), I composed my field notes the same day as the lesson. These field notes were then organized into analytic memos that track emerging categories and rich points (Agar, 1991).

The decision to engage in non-participant observation was undertaken after some reflection on the goals of the project. First, it was important to observe the actual content of the BIOME curriculum, rather than just lesson plans or summaries. The difference between a curriculum "on paper" and in practice can be large. With this in mind, I chose not to consult regularly with instructors or mentors about the content of the program or to spend significant time analyzing the syllabus before my observations. At each moment of the course, I chronicled events as they were actually unfolding, even when they did not appear to go according to the apparent plan for the day. For example, when students were directed to engage in table talk about a particular topic, they sometimes decided to wrap up an earlier discussion first, or spend only minimal time on the assigned topic. Relative to my experience as a teacher of undergraduates, participants were highly receptive to the directions given by the instructor, but they also complied in their own ways and at their own pace. Second, it was important to me to spend as much time gathering real-time information about events as they unfolded. Had I taken on an active role in the program as a participant observer, for example, by working as teaching assistant, my creation of such a detailed record would likely have been interrupted by my duties. Moreover, the technique of participant observation requires a role in which the researcher will be ratified as a legitimate participant, and such a role did not seem available in this context. While I have served in the roles of both teacher and academic

advisor, my limited knowledge of the biology program would have deprived students of a valuable source of information had someone more situationally competent filled the role.

The choices that led me to examine this particular class were more or less arbitrary in that I did not select the program out of any preexisting interest or knowledge. Within the program there are three sections that meet each week, and the section I selected to observe was based entirely on my schedule, and was not known to the instructor before the first day of class. Seats were assigned prior to my arrival and the instructor could not have predicted that I would choose to study this small group, thus my selection of the small group at Table 3 was effectively independent of other factors of the research design, and as far as I am aware, impartial.

It is worth noting that few program evaluations involve the collection of such extensive ethnographic data - attending every session of a 10-week program is time intensive. However, in comparison to ethnographies, rather than to program evaluations, 10 hours of field observation is minimal. Aware of these disciplinary expectations, I chose to supplement my field observations with small-group interviews and a survey.

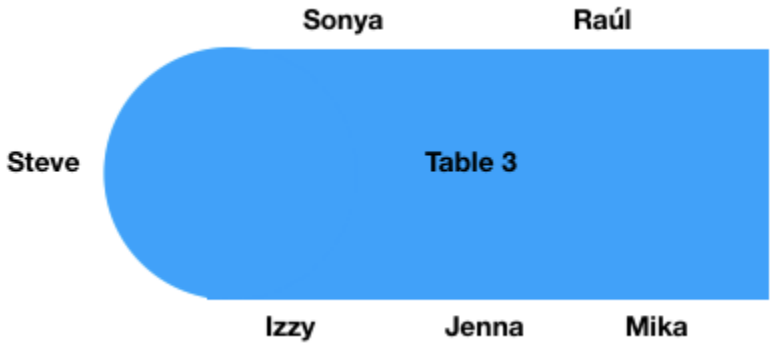


Figure 2.1: The seats chosen by first-year students and their mentor every day at Table 3.

2.3.2 Method 2: Semi-structured Group Interviews

Semi-structured group interviews were conducted during weeks 3, 7, and 11. Participants were recruited by the PI from the BIOME program across all three sections.[9] No course credit was offered for participation in focus groups. Each interview was begun with an explanation of the purpose of the discussion, a brief explanation of my role as an evaluator, and an opportunity for students to give or withdraw verbal consent to being audio-recorded. No students chose to withdraw from these sessions. The groups were composed of 5-6 first-year students and lasted approximately one hour each. A semi-structured discussion was conducted using an interview protocol as a prompt. The protocol focused on three key areas, 1) the positive and negative experiences of first-year students in their first quarter at university, 2) strategies used by first-year students in STEM classes, and 3) the value of the BIOME program specifically. The discussion of each of these topics was allowed to progress naturally until participants appeared to have exhausted their interest in each topic. Towards the end of each session, I prompted the students to compose individual written lists of positive, negative, and "in-between" features of BIOME, these artifacts serving as a stimulus to prompt any additional discussion of points that had not yet been mentioned. To conclude each session, participants were asked whether they would like to edit or revoke any prior statements. No participants opted to strike comments, and those who responded to this prompt used the opportunity to clarify earlier statements.

As Spradley has argued, interviews that resemble natural conversation are the easiest for participants to navigate (2016). According to Spradley (2016), semistructured interviews stake out a middle ground between highly structured interview questioning, which may break

[9] Trivial incentives such as homemade cookies were used for the initial focus groups, while gift cards were used for subsequent cognitive interviews.

discursive maxims such as reintroducing topics that have already been addressed, and unstructured conversation, which risks not arriving at answers to key research questions. The topic-ordering of the semistructured interview protocol moved deliberately from a more general "how are you doing in school so far?" to more specific frames about STEM classes and the focal program. In semistructured interviews, adding lines of inquiry *ad hoc* can make optimal use of participant knowledge by questioning participants who appear to know more about particular domains. Moreover, in these spontaneous moments of conversation I was able to engage in informal member checking of my developing ideas about the program. My interview protocol is supplied in Appendix I.

The timing of small-group interviews at multiple points throughout the quarter enabled the comparison of student points of view as they evolved over the duration of the quarter. Students in the first round of small-group interviews were very new to the university, while students in later interviews have established understandings and routines that help them navigate university life. This methodological choice reflects the theory that acculturation is a gradual process with multiple qualitatively distinct stages, and should be examined longitudinally within a community (Flick et al., 2004).

2.4 Qualitative Findings

In my effort to understand the way that BIOME functions, I recorded as much as I could observe of the talk and behavior of its members. In the process of transforming field notes into analytic memos, I sought to track the areas where the affordances of the course met the learning preferences of the students, as well as which affordances were not taken up, and which preferences were not met by the curriculum. As past course evaluations and member-

checking indicated, BIOME is a successful course attended by students who generally find it valuable. Thus, where rich points emerged in discourse of students, they tended to occur at moments of overlap between the affordances of the course and student preferences. Beginning with the welcome email, the social benefits of the course were emphasized:

> Hi BIOMErs! Welcome to MCDB 11/BIOME I where you will *meet other biology first-years*, gain an upper-div biology *mentor*, and learn all about UCSB and the Biology Major. We're excited to *meet* you all! (Sept 27, emphasis mine)

Throughout the mentor handbook provided by the program coordinators, the goal of socializing students into the norms of college life also appears in different forms. For example, mentors are directed to help students to understand when they are taking up too much time during discussions, as well as encouraging them to enroll in group tutoring. Modeling is another way in which the importance of social relationships is underscored. "Watch this everyone," the instructor says as he places a call to a tardy mentor from his cell phone. When promoting the idea of attending office hours, the instructor reassures students that faculty are open to relationships with them, "You can talk to professors about anything," he says. After class, a student mentions feeling trepidation about a course, to which the graduate student TA responds "My roommate TAs that class," and offers her direct help, "just send me an email if you need anything from her." As these examples illustrate, from the first moments of the course, college is represented as a network of persons endowed with and seeking knowledge. One maxim, repeated by the instructor several times throughout the quarter, was that students should "never to walk into a class on the first day without knowing how that class operates." In a presentation entitled "Crystal Ball and Time Machine", former pre-biology majors explain, via a slide show, what they wish they had known or done as first-years. Social relationships are the key to both the ontology and epistemology of the program:

university is group of people whose knowledge of both the rare and the mundane is the pupil's primary resource.

For their part, students express their preference for forming and regulating social relations in the university setting. On the first day of class, the instructor asks what students' goals are in taking the course: "I want to find study buddies," replies one student, prompting the instructor to request show of hands of who is in various required biology classes. During the first class, students are encouraged to exchange phone numbers with their mentor and group members. The communication threads some of these groups create last all quarter, updated frequently with questions for the group and news of success. "I'll see you in class tomorrow" a young woman says to a new acquaintance on her way out of class on the first day.

In the coming weeks, students appropriate many of the social goals of the mentors and instructor. In these moments, first-year students' discourses often harmonized with the affordances of the course, with some students requesting help and others acting in the role of counselors. When Mika expresses that she is having problems structuring her studying, the mentor begins by asking her whether she followed the recommended course of action: "Have you been going to CLAS" Steve asks. "It doesn't help," Mika responds, softly. The first-years jump in to suggest that she switch sections of CLAS and go to a different tutor. They compare notes on the different tutors. "Try the drop-in hours" Steve advises. "Have you gone to office hours?" he asks. "It's scary!" Mika protests. Sonya jumps in excitedly, "Go! Go! Go! It's not bad!" Through extended ethnographic study of a small groups of participants, I was able to observe the authentic uptake of ideas presented by the instructors and mentors by the students. On week 3 of the quarter (Oct 11th), for example, the instructor narrates a short

account of his own experience as a member of a tight-knit study group in college, concluding "There was solidarity in doing the work together." A month later (Nov. 8), Raúl fondly describes the peer study group he has formed, concluding with a shy smile "It's cool to struggle together," echoing the instructor's earlier phrase.

Figure 2.2. Sonya explaining her study habits, with great excitement. Artist's impression by Britta Young.

Other topics of discussion that merged the social and academic included how to learn useful information about future professors from other students, the optimal social relationships for members of a study group, and decorum for contacting faculty. Interestingly, while members of the BIOME community clearly oriented to a wide variety of social and academic practices, at no point was this category of habits given a name or

distinguished from other "strategies" used by students. On the second week of the course, a chemistry professor was invited to speak about how to prepare for chemistry assessments. "What is a strategy for first year students to go to office hours?" the instructor asked her, in a word choice that initially struck me as odd. As revealed by the subsequent conversation, the instructor did not mean to imply that going to office hours required some special trick, and his guest did not take him to mean this either. Rather, the instructor had adopted the frame that interactions between students and faculty were to be undertaken strategically, analogous to other elements of engagement with academic work such as "study strategies" or "testing strategies." The behaviors discussed by students were implicitly conceived as repeatable routines, rather than as solutions to isolated issues. Strategies were constructed as temporally extended orientations and patterns, formulas for success.

Observations such as these led me to center on a rich point (Agar, 1991) in first-year acculturation to the university: academic habits that involve a social component. Prior to engaging in the BIOME project, I was unaware of any literature on the issue of social academic skills (e.g. Tinto, 1975, 1987). That is, this was not among my sensitizing concepts (Charmaz, 2006), but emerged through the process of making field notes and memos in a grounded theory framework.

Simultaneously, some of the evidence I have collected also points to the other area of an outcome space (Wilson, 2004, p.67) implied by contrast with social academic skills - individual academic skills. Ideas about how to effectively schedule and structure one's individual study time also circulate in the program community. The program provided a place to ask questions not only about how to do one's work most effectively, but also when to do it, and how much to do. Directions in how to study – proposed by both students and

instructors – included the use of timers, noting missed problems, and attempting to identify target concepts rather than superficial features of problems. The question of when to study was addressed in a variety of ways, including a week-long exercise in logging how students' time was actually spent, and then comparing this to study goals. The proper amount of studying was frequently an explicit or implicit topic of conversation, brought up in talk about routines, planned study breaks, and well-being. As in the case of social academic habits, students requested explicit guidance in these areas throughout the duration of the course.

Individual academic habits were involved in a dialectical relation with the social habits. Each week first-year students were instructed to individually work a large number of practice chemistry problems. In practice, the accountability system for this individual academic work entailed a reporting to the mentor of how much individual work each person at the table had done. Each mentor then averaged the number of problems for the table and wrote this average on the board for the entire class to see. When advice about individual academic practices was offered, such as "Make use of old exams and practice exams, go back to notes, to the book, the ones you got wrong, find out where to focus" (Week 10), it was often in the context of a deliberate polling of the mentors by the instructor, underscoring the extent to which other people hold the informational keys to individual academic success. On several occasions, I observed conversations about individual study skills in which the mentor of Table 3 effectively marshalled his epistemic authority in discussion with students about individual study habits. Simultaneously, the instructor of the course rarely issued any categorical statements about optimal individual academic practices, allowing these ideas to emerge from dialogue with mentors and from course readings, which were then critically discussed.

The cumulative impression given by these classroom observations, with particular attention paid to student practices, was that a large number of qualitatively distinct behaviors were recommended to and habitually taken up by students. A great many of these practices were new to students, both because of putative differences between secondary and university environments, and because of the imposition of higher standards of academic performance necessitated new approaches. Some of these new academic habits were easily acquired and maintained, while students struggled with others. A few students clearly had many, highly developed academic habits. Sonya, a student at Table 3, cheerfully related her routines on several occasions, which included using timers to study, scheduling breaks during which she would listen to classical music, personalized quotas for daily chemistry practices problems, and more. While each day of the course focused on some particular topic - often a bundled subset of academic practices - the accumulation of these topics amounted to an implicit imperative: students needed to *complexify* their secondary school academic habits. I began to call this "academic habit complexity" in my analytic memos about the course.

In addition, I also identified a number of other potential constructs for further investigation. These included the constructs of "social skills", the construct of a "new (academic) identity", motivation, prior academic achievement or readiness, and grit (Duckworth et al., 2007). Participants showed signs of interest in habit complexity, social skills, academic identity, and motivation but not grit or prior achievement. For example, the concept of "grit" presented in the curriculum, was explicitly rejected by the students as "useless" for practical purposes. Among the identified constructs in which participants expressed interest, only habit complexity appeared to be a semiotic rich point. The other identified constructs were mentioned once or twice by participants but were rarely developed

through further discourse. By contrast, academic habit complexity was an organizing concept that characterized at least one rich discussion each week: to students, the topic was constructed as important, non-obvious, and of general interest.

2.5 Discussion:

2.5.1 Theoretical Considerations in Construct Selection

The criteria by which constructs should be selected for study and measurement is undoubtedly among the thorniest issues in social science. The process of construct selection is sometimes portrayed as a metatheoretical issue, either as a matter of researcher intuition or simply omitted from accounts of the research process altogether. All investigation is theory-laden, the argument runs, so we cannot get to a pre-theoretical position from which to adjudicate which theories to employ. Yet, as Thurstone acknowledged in his 1929 book *The Measurement of Attitude*, selecting one or more constructs is a logically necessary step for empirical research designs:

> The first restriction on the problem of measuring attitudes is to specify an attitude variable and to limit the measurement to that. . . . This restriction on the problem of measuring attitudes is necessary in the very nature of measurement. It is taken for granted in all ordinary measurement, and it must be clear that it applies also to measurement in a field in which the multidimensional characteristics have not yet been so clearly isolated (p.11).

In practice it is virtually impossible to study more than a handful of constructs at a time, meaning that choices have to be made even in the best of circumstances. Worse, the chronology of construct selection in the social sciences often belies any claims to inductive reasoning. Proposals for research, including grants and institutional review board approval, encourage or require that constructs be selected in advance of any data collection. Unless the

proposal is to continue research with the same population, contact with participants at this stage is typically minimal or informal. For many projects, this means that construct selection is undertaken at the moment when no current data are available about the focal population. Statistically-driven projects frequently select all constructs and proceed through the instrumentation phase prior to collecting any data. The entrenchment of disciplinary divides and recurrent flare-ups of the science wars have perhaps left us less willing than ever to ask: when the process of construct selection really is inductive, how is it done?

Historically, the quantitative and qualitative traditions have each treated inductive construct selection in their own way. Construct selection in the qualitative tradition is typically considered *before and during* data collection. The question of how to focus the analysis has received serious philosophical treatment, particularly within cultural anthropology (Moore & Sanders, 2014). The early days of anthropology saw many valiant attempts to make comprehensive records of "The Cultural Practices of the X People", until both academic specialization and the obvious impossibility of fully delivering on such promissory titles took their toll.[10] More focused ethnographies are now the norm, and books are now more likely to focus on either the life stories of a small group of participants or just one aspect of the cultural traditions of a culturally distinct group. The practical and theoretical benefits of narrowing the scope of analysis have made of construct selection more than a necessary methodological step - the logic of construct selection is also part of the heritage of qualitative research. Perhaps the hallmark strategy for inductive construct selection within anthropology has been to select constructs that are of interest to participants

[10] E.g. Anahuac: or, Mexico and the Mexicans, Ancient and Modern (Tyler), The Andaman Islanders (Radcliffe-Brown), The Nuer (Evans-Pritchard), Balinese Character: A Photographic Analysis (Bateson & Mead)

(Geertz, 1973, p.453). When participants possess explicit models of some phenomenon, these deserve consideration. As Levi Strauss argued, "these models might prove to be accurate or, at least, to provide some insight into the structure of the phenomena; after all, each culture has its own theoreticians whose contributions deserve the same attention as that which the anthropologist gives to colleagues" (2008, p.282). Allowing participants to guide the researcher in the construct selection process is one protection against the erroneous imposition of the researcher's preferred constructs - and the models they compose - on a situation which we do not yet understand. In addition to following participants' own interests, qualitative research often selects constructs that are semiotically dense within a cultural context, so-called "rich points" (Agar, 1991). Doing so requires a higher level of interpretative abstraction than the strategy of following participants' interests, since it involves the careful sifting of evidence from situated discourse and practice. It is also logically possible that a rich point may not be acknowledged or explicitly discussed by participants, while being nonetheless apparent from their behavior - taboos representing the extreme case. Once a semiotically dense domain has been identified by the researcher, the appropriateness of this interpretation of things can be verified by checking with participants (Brenner, 2006, p.268), who may accept or reject it.[11] Both of these strategies are forms of inductive reasoning about which constructs to select for analysis, and both are common in qualitative research. Both strategies proceed by naming observed patterns that appear to occur in streams of information, and both treat participants as one important source of information about the question of which constructs matter.

[11] Not all anthropologists are this fastidious, of course, but even the data generated by such attempts when they fail is highly useful to the anthropologist's task of theory development.

In quantitative research, inductive construct selection is typically undertaken *after* data collection has occurred, as part of the process of "model selection." The data are taken as given. Since model selection happens after data collection this process is eliminative in nature: models are selected for parsimony by excluding variables that do not explain a significant amount of variance in the outcome. In his classic textbook, the relentlessly quantitative Eric Hanushek argues that variable selection is "perhaps the most important topic in this book" even as he explains that the actual process departs considerably from statistical methods (Hanushek & Jackson, 1977). "The specification of behavioral models relies upon the purposes of the model, the available theories of behavior, the past empirical forays into an area, and the embodied wisdom and hunches of the researcher and associates," Hanushek and his coauthor write, but "It is not feasible for us to discuss how these elements are accumulated or combined" (1977, p.80). Such deferrals, while frustrating to the reader, are typical of quantitative approaches to construct selection: the issue is all-important but somehow beyond the scope of the tutorial. In practice, the problem is approached from a different angle: signs of problems with the variable selection process are sometimes sought, again after data collection. Checking for signs of omitted variable bias is one way of addressing problems with construct selection. Omitted variable bias is discovered via distortion in the statistical model, evidence of which can sometimes be detected via analysis of the distribution of residuals. No sooner has the call to find omitted variables been trumpeted, however, than the hunt takes a rapid turn for the metaphysical: omitted variable bias in the predictors is inevitable, as acknowledged in virtually every statistics and econometrics textbook. In practice, the most common approach for handling omitted variable bias is the inclusion of control variables, although there is no mathematical

justification for this method (Clarke, 2005). In addition to checking for omitted variables, it is typical in quantitative research to compute some proportion of the variance in the outcome variable that has been explained by the model (e.g. 20%), and to state that additional constructs might have helped bridge the gap between this and a more comprehensive explanation of variance. Estimation of this gap also alerts the reader that some relevant constructs are likely missing from the model of the process. While these diagnostic strategies are of course helpful, they are neither necessary nor sufficient for construct selection.

To summarize, the qualitative research tradition is endowed with methods to discover and name new constructs. Qualitative researchers accomplish this feat as a standard element of their analysis. They issue the frequent reminder that, where student attitudes and behavior are concerned, "the old maps are obsolete" (Flacks & Thomas, 2007). By contrast, statistical research uses strategies that are delimitative rather than denominative - they prune and shape models and constructs instead of discovering and naming them. There are means of cutting away constructs that do not contribute to the overall analysis and for highlighting the gap between what the model can explain and what it hopes to explain. Both qualitative and quantitative methods contribute to construct selection, since selection inescapably involves both discovery and choice of constructs. We can think of the combination of denominative and delimitative methods for construct selection as a Baconian inductive process in which generalizations - named patterns - are tested and reduced via subsequent checks and challenges.

In addition to these inductive methods, both qualitative and quantitative methods have also historically employed non-inductive approaches to construct selection, often rhetorically motivated by "theory." Such approaches are a negative image of the classic anthropological

method of construction selection above, since instead of being guided by the ideas of participants they are guided by the ideas of non-participants. The constructs selected are the "theoretical terms" in the model or models favored by the researcher. Comparative and experimental studies often employ the theoretical method of construct selection. In qualitative research, the researcher may locate communities in which certain processes are thought to be occurring in order to compare them to communities which are not undergoing these processes, implicitly privileging prior theory. In experimental quantitative research, construct selection is sometimes motivated by the need to test multiple theories. For example, in order to test whether self-determination theory is a better model of behavior than interest-enhancing theory, all the constructs from these models may be selected for inclusion in a single study (Jang, 2008).

One realist criterion for determining whether the "theory" is appropriate is to ask whether it describes the thought and behavior of the selected population. In the social sciences, evidence about whether these are the correct constructs for the focal population often consist of model fit, using confirmatory factor analysis or structural equation modeling, but this misses the mark for at least three reasons. First, these methods do not directly address whether the constructs are appropriate for the population, since person fit is not typically estimated.[12] This means that many studies make no claims about what proportion of the population is well-described by the model. Second, false positives in construct selection - including the extreme scenario of the inclusion of empty constructs that lack real referents - are trivially easy to generate in empirical studies (Maul, 2017). Just as the model may not include some critical construct, constructs that are represented in the model may not actually

[12] Methods of integrating fit statistics into covariance structure models have been proposed (e.g. Reise & Widaman, 1999), but to my knowledge these are rarely employed.

be applicable to the target population. Third, multiple plausible models including different constructs frequently fit the data equally well. While in many studies, estimates from several plausible models are published side by side, the fact that these models fit the data equally well means that model fit is not a sufficient condition to declare construct selection a success.[13] All three of these major problems can occur without raising red flags in the standard practice of statistical analysis.

Where "theory" postulates previously unobserved constructs, this is usually by analogy. For example, the "double consciousness" of African Americans (Du Bois, 1994) may serve as a useful point of departure for theorizing the triple consciousness of Black women under patriarchy (Welang, 2018) or, more speculatively, the double consciousness of people living with profound physical disabilities. These are forms of “abduction” in the comparative, lateral sense of metaphorical thinking postulated by Bateson (1979). However, theorizing a new construct in this way is at best a "sensitizing" (Charmaz, 2006) exercise with which to begin inquiry, since there is no guarantee that such constructs will be operative in the focal population. At worst, it may prejudice the researcher to emphasize similarities with some previously-theorized construct. To summarize, selecting constructs based on theory entails the adoption of a confirmatory mode of enquiry - constructs are presumed to be applicable to members of the target population, even when no evidence for this tacit claim has been collected.

I argue that these methodological criteria for construct selection amount to what Ludwik Fleck called active or passive elements in a thought style (Fleck, 2012). That is, these criteria condition perceptions of incoming information, supply standard questions to be answered,

[13] In the statistics literature, this family of problems is known as "model dependency", since the estimates produced depend on the arbitrary selection of a model from among several that fit the data equally well.

and furnish stopping rules for inquiry. Mixed methods research presents an opportunity to combine criteria (active and passive elements) from different thought collectives. In the following section, I propose a constellation of criteria for construct selection in a context of low methodological pre-specification.

2.5.2 Selection of a Focal Construct

Ethnographic methods were employed to inductively select a construct for further study. Direct non-participant observation of student talk and behavior allowed me to make some initial judgments about student interests, narratives, and plans. In the group interviews, I guided the conversation towards these areas to judge the reactions of additional students as a form of member-checking and probing. Although several constructs were identified using these methods, I selected just one for the development of a measure. This process of construct selections merits further reflection as a case study of a discovery-based workflow. Several constructs were identified via analysis of field notes. Some of these constructs were directly referenced by students, while other constructs were probably more noticeable to me than to participants. Constructs identified by students included "social skills", "pushing oneself", and "getting away from one's background." Additional constructs were introduced by the professor of the course, and some were taken up by students during the course, while others were not. For example, when presented with the concept of "grit" (Duckworth, 2007), students ridiculed it as vague, static, and unhelpful - there was no talk about the concept afterward for the rest of the term, the case was closed. In my notes, I also identified constructs that may have been operative but of which students may not have been explicitly aware. I have several years of classroom teaching experience and, like any teacher, have my

own theories about students. I tried to the best of my ability to bracket out these preconceptions in my analysis of the unfolding ethnographic situation, a task accomplished by reminding myself of the potential of each classroom and group of students to develop its own unique dynamics.

Having identified several constructs that might be at play in the situation, I turned to the difficult task of applying selection criteria. Possible criteria included the interests of non-participants (program leaders, myself), the interests of participants (students), purported ability to explain variance in an outcome, or parsimony. The anthropological criterion for construct selection - that is, constructs should be selected for analysis on the basis of participant interest (Geertz, 1973) and semiotic richness (Agar, 1991) - seemed most germane to this form of classroom research. While the constructs introduced by the professor of the course may indeed have been helpful for students if the latter had chosen to appropriate them, my study was focused on the students rather than the instructors. For measurement of attitudes and behaviors to be successful, participants must share specific patterns of thought and action - a "marriage satisfaction" instrument does not measure anything if administered to bachelors, even if they answer all the questions. From a measurement perspective, there seemed little reason to believe that a construct like "grit" had reached the threshold of intersubjective diffusion in the student population necessary for measurement. There were a number of possible outcomes that a construct might have been selected to predict: academic performance, retention in a major, satisfaction with the university experience overall, strength of the desire to pursue a particular career, or more distal outcomes. Choosing a construct based on its purported ability to predict one of these outcomes would have involved more guesswork than "hypothesis." The predictors of success

were thus treated as truly unknown. Likewise, selecting constructs on the basis of parsimony requires either some outcome towards which to optimize or the atheoretical application of a dimension-reducing procedure such as PCA.[14] Again, these approaches appear to presuppose far too great a degree of knowledge about the population and which constructs are relevant to its attitudes and behaviors. Parsimony may be a virtue in general, but applied too early in theory development it may ultimately reduce the possibility of arriving at the optimal model - a classic case of premature convergence, representing an unfavorable explore-exploit tradeoff (Axelrod, 2000).

To make explicit the process of applying selection criteria to identified constructs, I have provided a table showing how these criteria are applied some identified constructs (Table 2.1). In this table, identified constructs are given a construct label and then judged by construct selection criteria of participant concern, stakeholder and/or researcher concern, semiotic richness, and plausible connection to relevant outcomes. This table summarizes available qualitative evidence about the participants, and is not meant as a final judgment on the constructs themselves. Perhaps the most intriguing combination of these criteria were constructs that were named by participants but never elaborated upon. This discursive patterning may be due either to a high level of presupposition, topic sensitivity (e.g. drug abuse), or the emptiness of the construct (e.g. "the right stuff" among test pilots).[15] In future research with this population these are constructs that merit further exploration, perhaps through individual interviews. Prescriptively, not all of these criteria need to be met in order

[14] As a thought experiment, imagine the use of PCA on the characteristics of a population of a sports team to derive an overall "fitness component" for each team member, combining height, weight, running speed, and so on. Will this fitness component be the best predictor of the variance in overall performance in the sport? Only if skill in the sport matters little. Knowledge of the nature of the activity is key in selecting plausible constructs and models.

[15] Chuck Yeager, one of the best test pilots in the space program, explained to Tom Wolfe that "the right stuff" was really knowledge of the aircraft plus luck - not a personality characteristic.

for a construct to be selected for further analysis. However, some combinations of these criteria should heavily tip the scale for against the inclusion of some construct. A construct which is of interest to stakeholders and researchers but not to participants and which is not a rich point seems unsuitable for measurement - indeed, participants may not even comprehend such a construct without having been initiated into some theory. A construct that meets all other criteria but is not of interest to researchers or stakeholders still be considered for measurement in case our theory is wrong. I suggest that any inductive process of construct selection should be amenable to the creation of such a table, and should attempt to make an argument for the selection in terms of criteria similar to these - my list of criteria is not exhaustive and a more complete enumeration would be welcome. Critically, I wish to point out that construct selection is a normative decision that can be modeled as a decision between constellations of these criteria and others.

Of the identified constructs, adequate information for only one existed to meet these criteria in a compelling manner. This was the construct of academic habit complexity. Talk about this topic was facilitated by the structure of the intervention, but was also fully appropriated by the students. The topic was discussed colorfully both spontaneously and when elicited by the instructor. Academic habits were a rich point that seemed to require endless re-examination and qualification. When a participant shared their experiences with some habit or routine, other participants listened attentively and contributed to the discussion of the topic. Participants were highly interested in the question of what combination of academic habits would be most likely to lead to success in the difficult STEM coursework they faced immediately upon beginning their studies at the university. Given the social and institutional structure of the university, this was a plausible connection between behavior and

outcome. That is, given a certain level of prior knowledge, grades in STEM coursework at this stage appear primarily work-related, rather than primarily a product of students' relationships with instructors or other factors. It is thus plausible to suppose that academic habits might be causally connected to outcomes.

The feature of academic habit complexity that requires the most justification in the construct selection process is the choice of "complexity" rather than some other dimension of habit, such as "strength." Here complexity refers to the number and variation of academic habits, while strength (Verplanklen & Aarts, 1999) refers to the extent to the amount of time and effort dedicated to each. The choice of habit complexity as the first construct to measure, as opposed to habit strength, was motivated by the semiotic density (Corrington, 2000; Yanushkevich, 2014) of complexity. In short, participants did not seem as interested in the idea that they needed to work harder and longer. A presupposition of several discussions about academic habits was that habit intensity was insufficient for success. However, these two dimensions of academic habits are not mutually exclusive - participants might have habits that are both intense and complex, complex but not intense, intense but not complex, or neither intense nor complex.[16] Because of this relationship, habit strength is a logical target for future measurement.

Prior theory was not consulted for the selection of this construct. Indeed, I am unaware of any prior theory of habit complexity to which I might have referred. This state of affairs underscores the importance of inductive, rather than theory-driven, construct selection. True, I was aware of the culturally-constructed notion of "habit" and some of the academic

[16] Future iterations of the instrument have attempted to model habit intensity as well since this is theoretically a dimension of the larger construct of academic habits. See Chapter 4 ("Criterion") for an exploration of attempts to use existing measures as a proxy for habit strength.

literature about habit, but if I had not been aware of habit prior to my investigation, the talk and behavior of the students would shortly have taught me the importance of this concept. The academic literature on habits turns out to be focused on several adjacent concepts to habit complexity: habit strength, academic integration, and study skills (see Chapter 3). None of these are equivalent to the notion of habit complexity, and had I selected on any of these constructs too early, I may have missed the distinctive way in which the habits essential for success were being socially constructed in this population. The methodology of grounded theory is an invaluable companion for researchers engaging construct selection (Strauss & Corbin, 1994) - instances of construct selection which are not compatible with grounded theory should consider why their alternative is preferable.

Table 2.1

Some Constructs Identified and Criteria for Construct selection

Identified construct label	Participant concern	Stakeholder/ researcher concern	Semiotic density (rich point) among participants	Plausibly connected to relevant outcomes
Grit	No (rejected)	Yes	No	Yes
Habit complexity	Yes	Yes	Yes	Yes
Social skills	Yes	Some	Some	Unknown
New academic identity ("Getting away from one's background")	Yes	No	No	Unknown
Motivation ("pushing oneself")	Yes	Yes	Some	Yes
Prior academic achievement	No	Yes	No	Yes

The selection of the construct of academic habit complexity for measurement was thus motivated by extensive attention paid to this rich point by both students and instructors, in and out of class. While it would have been possible to select other constructs to investigate, academic habit complexity was chosen because of the interest paid by students to constructing behavioral formulas for academic success. I hoped to use an instrument to capture a dimension of student life that was never spoken of in explicit terms but that implicitly motivated much talk and behavior. Moreover, it seemed an open question whether participants who advocated for more complex formulas of academic habits and routines were correct in their belief that these would lead to academic success. Perhaps, after all, the best way to succeed in a tricky STEM course would be to simply spend more hours alone, engrossed in the textbook, or to remind oneself at regular intervals not to give up. By testing the idea that a complex formula of academic behaviors might be more effective than this, I was also testing a hypothesis which I hoped would be legible and actionable for students and instructors.

References

Agar, M. (1991). The biculture in bilingual. *Language in Society*, 20(2), 167-182.

Axelrod, R., & Cohen, M. D. (2000). *Harnessing complexity*. Basic books.

Bateson, G. (1979). *Mind and nature: A necessary unity*. New York, NY: E. P. Dutton.

Brenner, M. E. (2006). Interviewing in Educational Research. *Handbook of Complementary Methods in Education Research*, 2.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Sage.

Chen, H. T. (1996). A comprehensive typology for program evaluation. *Evaluation practice*, 17(2), 121-130.

Clarke, K. A. (2005). The phantom menace: Omitted variable bias in econometric research. *Conflict management and peace science*, 22(4), 341-352.

Corrington, R. S. (2000). *A semiotic theory of theology and philosophy*. Cambridge University Press.

Crede, E., & Borrego, M. (2013). From ethnography to items: A mixed methods approach to developing a survey to examine graduate engineering student retention. *Journal of Mixed Methods Research*, 7(1), 62–80.

Creswell, J. W., & Plano Clark, V. L. (2011). Designing and conducting mixed methods research (2nd ed.). Thousand Oaks, CA: Sage.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., ... & Weiner, S. S. (1980). *Toward reform of program evaluation* (p. 3). San Francisco: Jossey-Bass.

DePass, A., & Chubin, D. (Eds.) (2009). *Understanding interventions that encourage minorities to pursue research careers: Building a community of research and practice.* Bethesda, MD: American Society for Cell Biology.

Du Bois, W. E. B. *The Souls of Black Folk*. New York, Avenel, NJ: Gramercy Books; 1994.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology*, 92(6), 1087.

Emerson, R. M., Fretz, R. I., & Shaw, L. L. (2011). *Writing ethnographic fieldnotes*. University of Chicago Press.

Estrada, M., Burnett, M., Campbell, A. G., Campbell, P. B., Denetclaw, W. F., Gutiérrez, C. G., ... Zavala, M. (2016). Improving underrepresented minority student persistence in STEM. *CBE—Life Sciences Education*, 15(3), es5.

Flacks, R., & Thomas, S. L. (2007). 'Outsiders', Student Subcultures, and the Massification of Higher Education. In *Higher education: Handbook of theory and research* (pp. 181-218). Springer, Dordrecht.

Fleck, L. (2012). *Genesis and development of a scientific fact*. University of Chicago Press.

Flick, U., von Kardoff, E., & Steinke, I. (Eds.). (2004). *A Companion to Qualitative Research*. Sage.

Geertz, C. (1973). *The Interpretation of cultures*. Basic books.

Hanushek, E. A., Jackson, J. E., & Jackson, J. E. (1977). *Statistical methods for social scientists: Quantitative studies in social relations*. New York, NY: Academic Press.

Hitchcock, J., Sarkar, S., Nastasi, B., Burkholder, G., Varjas, K., & Jayasena, A. (2006). Validating culture- and gender- specific constructs: A mixed-method approach to advance assessment procedures in cross-cultural settings. *Journal of Applied School Psychology*, 22(2), 13–33.

Jang, H. (2008). Supporting students' motivation, engagement, and learning during an uninteresting activity. *Journal of Educational Psychology*, 100(4), 798.

Lévi-Strauss, C. (2008). *Structural Anthropology*. Basic Books.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, 15(2), 51-69.

Moore, H. L., & Sanders, T. (Eds.). (2013). *Anthropology in theory: issues in epistemology*. John Wiley & Sons.

Nastasi, B., Hitchcock, J., Sarkar, S., Burkholder, G., Varjas, K., & Jayasena, A. (2007). Mixed methods in intervention research: Theory to adaptation. *Journal of Mixed Methods Research*, 1(2), 164–182.

National Center for Education Statistics. (2005). Digest of education statistics. Retrieved July 9, 2020, from https://nces.ed.gov/programs/digest/ d05/tables/dt05_009.asp

Latour, B. (2012). *We have never been modern*. Harvard university press.

Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational researcher*, 33(2), 3-11.

Otero, V., Pollock, S., & Finkelstein, N. (2010). A physics department's role in preparing physics teachers: The Colorado Learning Assistant model. *American Journal of Physics*, 78(11), 1218–1224.

Schensul, J. J., & LeCompte, M. D. (2012). Ethnographer's toolkit. Rowman Altamira. Walnut Creek, CA.

Scriven, M. (1973). Goal-free evaluation. *School evaluation: The politics and process*, 319-328.

Solanki, S., McPartlan, P., Xu, D., & Sato, B. K. (2019). Success with EASE: Who benefits from a STEM learning community?. *PloS one*, 14(3).

Spradley, J. P. (2016). *The ethnographic interview*. Waveland Press.

Strauss, A., & Corbin, J. (1994). Grounded theory methodology. *Handbook of qualitative research*, 17, 273-85.

Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude*. Chicago: University of Chicago Press.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45, 89–125.

Tinto, V. (1987). *Leaving college: Rethinking the causes and cures of student attrition*. Chicago: University of Chicago Press.

Verplanken, B., & Aarts, H. (1999). Habit, attitude, and planned behaviour: is habit an empty construct or an interesting case of goal-directed automaticity?. *European review of social psychology*, 10(1), 101-134.

Welang, N. (2018). Triple consciousness: The reimagination of Black female identities in contemporary American culture. *Open Cultural Studies*, 2(1), 296-306.

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Routledge.

Wilton, M., Gonzalez-Niño, E., McPartlan, P., Terner, Z., Christoffersen, R. E., & Rothman, J. H. (2019). Improving Academic Performance, Belonging, and Retention through Increasing Structure of an Introductory Biology Course. *CBE—Life Sciences Education*, 18(4), ar53.

Xu, D., Solanki, S., McPartlan, P., & Sato, B. (2018). EASEing students into college: The impact of multidimensional support for underprepared students. *Educational Researcher*, 47(7), 435-450.

Yanushkevich, I. (2014). Semiotics of social memory in urban space: the case of Volgograd (Stalingrad). *International Journal of Cognitive Research in Science, Engineering and Education*, 2(1), 43-50.

Chapter 3

Construct Definition

3.1 Introduction

The label "academic habit complexity" denotes a novel construct in educational research. This chapter is dedicated to the definition of this construct. It begins with an explanation of the necessity of definition for measurement, with an emphasis on the motives of this decision from a sociocultural perspective. This is followed by a definition of the substantives "habit" and "complexity" along with the qualifier "academic." Having proffered a definition, the chapter concludes with a discussion of the ontology of the construct and a normative argument for the study of this construct in educational settings. Since the objective of defining this construct is to provide warrants for later measurement claims, aspects of the construct definition that imply affordances for measurement are examined throughout the chapter.

Defining constructs in social science is critical to enhancing the rigor of empirical research for five key reasons: 1) definitions allow constructs to be generalized to contexts outside research situations, 2) definitions set parameters on new instrumentations of the same construct, 3) confusion about definitions leads to uncertainty about the meaning and comparability of results (BiPM et al., 2012, §2.27), 4) definitions specify theoretical aspects of the construct that are not measured, and 5) definitions clarify the meanings of the numbers that result from measurement.[17] Failure to explicitly define constructs may lead to the

[17] For example, "extroversion" can only be truly generalized to non-research contexts if it is defined as something other than as a response pattern on an instrument of extroversion. One would also theoretically like to be able to create new measures of extroversion that are not identical to the existing instruments, but only a definition can provide explicit parameters for doing so. Additionally, the definition of extroversion will presumably include references to theoretical terms, such as traits and states, that are not explicitly mentioned in measures themselves. Finally, a numeric score on an extroversion measure is only meaningful if the score can be cashed out in the terms of a definition of extroversion.

presumption that construct definitions are identical with the instrumentation of the construct, that is, the fallacy of "operationalism" (Bridgman, 1927; McGrane, 2015). The five reasons given here are not necessities flowing from transcendent characteristics of measurement, such as the principles of conjoint measurement (Luce & Tukey, 1964), but are aspects of the sociocultural context of educational measurement. To wit, this disciplinary context features a proliferation of ad hoc constructs (e.g. learning styles, Cuevas, 2015), defined and enacted with immense variability, often with scant validity evidence.

The reasons for providing construct definitions inform the ideal method for doing so. As Richard Pring (2000) has argued, definitions in educational research should be more than stipulative or ostensive. Stipulative definitions set prescriptive criteria to establish the meaning of a term, such as "by a student, we mean someone who is currently enrolled at an accredited educational organization." While stipulative definitions are useful for constructing watertight arguments, they do not entail a strong commitment to the truth of the claims embodied in the definition - that is, in the above example we will likely be able to acknowledge instances in which a "student" is not enrolled (e.g. there has been a clerical error). Moreover, stipulative definitions can overreach, since they are contingent *a priori* truths (Kripke, 1980). Even if a person is obviously *not* a student, stipulative definitions bind the user to the claim that anyone who meets the criteria must be a student (e.g. a person who has just emphatically quit school but remains enrolled). These dynamics of stipulative definition suggest that they may be most useful for creating genuine definitions when combined with efforts at falsification. Ostensive definitions point to examples of the phenomenon being defined, stating or implying that to be an example of X is to be similar to some known example *j* which is a member of set X, that is: $(j \in X \,\&\, j \simeq k) \rightarrow k \in X)$. In the

above example, showing an image of person sitting at a school desk might would be an ostensive definition of a student, as would recounting a narrative involving this person's daily activities. Ostensive definitions usually make recourse to other forms of definition in order to convey the intent of the user, since the object being defined is usually similar or different with respect to different qualia. Ostensive definitions cry out for decision rules for determining which cases belong to the focal category. Stipulative definitions offer one remedy for this problem with ostensive definitions, but ostensive definitions do not shore up any of the aforementioned weaknesses of stipulation. Pring (2000) suggests that two additional approaches to definitions are warranted: what we might call a usage-based approach and a conceptual approach. The usage-based approach, of which Wittgenstein's *Philosophical Investigations* (2001) could be considered the paradigmatic example, involves attending to the usage of a term in real situations. To formulate a definition of student-hood, we would gather and study instances in which the concept of student was invoked, leading to a definition that is a summary of these uses. Using this method, we may quickly find that "student" is a social status rather than an administrative category, that many supposed students do little studying, and so forth. The major liability of this approach, particularly from a philosophical perspective, is that some common uses of a term entail both internal and relative contradictions, rendering abstract summary of use impossible. We may discover that there is in fact some principled cultural dispute about the meaning of student-hood. When investigations yield such antinomies, we are forced to turn to the fourth approach to definition explicated by Pring (2000) - what might be labeled a *conceptual* approach. This approach is most akin to traditional philosophical definition because it requires us to "think of the different forms of understanding that are brought together" under the construct label

(Pring, 2000, p.11). These divergent forms of understanding the construct often turn out to represent contested values. For example, a philosopher who values autodidaxy may take great pains to define "student" in such a way that attending school does not figure in the definition. The conceptual approach involves drawing out implicit assumptions about the applications of the term and elaborating on traditions of thought. In this chapter, usage-based and conceptual definitions are privileged as the primary definitional strategy.

3.2 Habits

The notion that there are behaviors and habits that contribute positively to academic learning is an ancient one that has survived to the present day, albeit with some modifications. Aristotle emphasized the extent to which behavior and habits produce mental states, personality, and character:

> "[A] state [of character] results from [the repetition] of similar activities. That is why we must perform the right activities, since differences in these imply corresponding differences in the states. It is not unimportant, then, to acquire one sort of habit or another, right from youth. On the contrary, it is very important, indeed all important" (*Nicomachean Ethics*, Book II, 1103b.21-25).

The notion of repeated similar activities contains the core criteria for the concept of habit: 1) there must be an identifiable "activity", 2) this activity is repeatable, and 3) there must be an identifiable category of activities such that an activity can be considered an instance of it. It follows from the first criterion that subjectively-caused behavior that does not count as an activity cannot be a habit. Feeling melancholic at sunset may be a recurrent event in one's life, but it is not an "activity" and thus cannot be a habit. From the second criterion, it follows that unique activities cannot form habits. One cannot make a habit of graduating from high school or navigating any non-recurring situation. From the third criterion, it follows that

activities that fall into different categories cannot constitute a single, unified habit. One cannot have a single habit composed of both exercising and writing books - these are two separate habits - but one can have a habit of daily writing that includes both writing books and articles. To these criteria, we might add that habits consist only of agentive activity, that is activity that is within the volitional ambit of the individual. One does not have a habit of breathing. To summarize, when activities are members of a category of other similar, repeatable, agentive behaviors, then these activities are potential habits.

These types of activities hold a special place for Aristotle because they exert a causal power to shape character, and ultimately, foster virtue (MacIntyre, 2013). Non-Western intellectual traditions such as the Neo-Confucianism (Liu, 2018) likewise point up the ways in which habits set one to act with appropriateness (*yi*). Since habits are such powerful and cumulative shapers of human life, it is very important that even the small habits we choose arc towards the good. This argument, present in much philosophical discourse about habits, highlights another key feature of habits: namely, that habits are unlikely to be treated as morally neutral. When habits become an object of public discussion or measurement, they are typically invested with normative significance that is obvious to all members of the social group or measurement situation. This moral valence can pose deep difficulties for measurement insofar as it raises the specter of social desirability bias (Krumpal, 2013) in nearly all self-reports of habits. Since the moral valence of habit is deeply bound up in their philosophical character and sociocultural context, there seems little point in attempting to create self-report measures that are immune to the threat of social desirability bias. To measure only habits that are morally trivial activities would risk avoiding inquiry into the most socioculturally important questions. Because habits are so morally important, we will

probably always want to know the prevalence of cigarette smoking, reading, exercise, and many other habits.

One potential measurement solution to this dilemma is worth consideration: avoiding self-report in favor of other-report, that is, creating measurement situations in which participants report on someone else's habitual activities, rather than their own. In the case of habits, other-report involves a tradeoff between bias and resolution. Individuals may show bias about their own behavior. Resolution in this case refers to the degree of detail with which an observer can report on what is observed. By analogy to a microscope, telescope, or visual display, low resolution observations can detect large phenomena but may miss small phenomena. In terms of habits, an other-report by a teacher may be able to detect how many days per week a student turns in his homework, but this method will not have high enough resolution to detect how often he solved ungraded practice problems - the resolution of the measurement situation is too low to detect this latter detail. For habits, the highest measurement resolution is likely to come from self-report, since individuals are in general more aware of their own actions than of the actions of others. Other-report may be valuable for the measurement of easily-observable actions for which there is high social-desirability bias, such as smoking.

A second potential approach to solving the inherent social-desirability bias in habit measurement is to present queries about habits as queries about individual behaviors. This approach minimizes the criterion of repeatability of habits in favor of reporting isolated, morally trivial behaviors. For example, rather than asking about whether the participant completes practice calculus problems habitually, the participant can be asked whether she has completed any practice calculus problems in a recent interval of time, such as two days. This

formulation of the query allows for high resolution since it is a self-report, while treating the potential moral implications of habit more cautiously. That is, while smoking a single cigarette is arguably morally trivial, a smoking habit is socially undesirable. By avoiding the implication that individual behaviors index repeated behaviors, self-report measures can avoid triggering a morally defensive response. This method of measurement follows from the nature of habit as a morally charged, subjectively-caused activity.

3.3 Complexity

Complexity, the philosopher Edgar Morin states, is "a fabric (*complexus*: that which is woven together) of inseparably associated heterogeneous constituents: it poses the paradox of the one and the multiple" (Morin, 2015, p.21).[18] The constituents of complex constructs are unified by the emergent function of the whole. The modern era has been characterized by repeated attempts at hyper-simplification of physics and biology that have revealed greater and greater complexity. Given the complexity of the natural world, "anthropo-social phenomena cannot be expected to obey principles of intelligibility less complex than those required for natural phenomena," Morin argues, "We must confront anthropo-social complexity and no longer dissolve it or hide it" (Morin, 2015, p.22). By this standard, models of human behavior that do not account for complexity risk obscuring the very phenomena they seek to elucidate.

Accounting for complexity at the definitional phase of the measurement process involves describing the internal relationship of the construct's components and determining whether

[18] All quotes from Edgar Morin appear originally in French. Translations are mine.

these relations possesses the hallmarks of complexity. Ethnographic study of academic habits reveals that these habits are diverse, synergistic, and subject to selection pressures.

To begin, academic habits are an inherently diverse set of activities. They include activities that are individual and social, easy and challenging, general and specific. Like eating a nutritionally-balanced diet, good academic habits require the acquisition of fairly sophisticated knowledge and routines, employed with appropriate variation and at the right times. These habits are *diverse* rather than simply *variable* because they differ in type: working practice problems is a different kind of activity than creating a study schedule. Diversity itself traces a variable continuum often defined in terms of the number of types and the distribution of cases across those types. The similarity or distance between types can also be considered as part of diversity. Taken together, these criteria have been formalized in indices of entropy. A highly entropic construct would involve many types, with considerable variation between these types, and uneven distributions of cases among these types. "Music" would be an example of such a highly entropic construct. Academic habits seem to fall somewhere between the most entropic and least entropic constructs: there are several types of academic habits and with uneven distributions of cases among these types, to be sure. Yet, these habits are not innumerable - a reality hinted at by the fact that creative individuals are not constantly inventing new ways to study for chemistry.[19] Moreover, while these habits certainly vary in type, the types are not cosmically distant from one another since they are directed towards academic success. In other words, the qualifier "academic" rescues the construct from the high entropy of, for example, "habits of undergraduates."

[19] Evidence for this latter point is presented in the "Validation" chapter.

In addition to being diverse, academic habits are synergistic. For example, attending office hours after completing practice problems is a fundamentally different form of engagement than attending office without having done prior work. Isolated activities provide little marginal benefit. Tipping points occur when multiple combined strategies coalesce into an effective routine. For example, a student may read the chapter, try the practice problems, read the chapter again, create a personal quiz, then meet with her study group - all before moving on to new material. Different routines, defined here as bundles of habits, create contours in the landscape of habits, with peaks of effectiveness and fallow lowlands. Synergies may also occur at the inter-personal level when students positively influence each other's habits by word or example. Indeed, the focal program in this study relies on such inter-personal dynamics through its undergraduate mentors. These inter-personal relations are synergistic to the extent that engaging in some of academic habits, such as group study, may lower the difficulty of engaging in other habits.

Students are able to monitor the effectiveness of their own learning strategies (Hacker et al., 2009, Proust, 2013). Although some students appear more adept than others at this form of metacognition, all college students are capable of this to a certain extent. The ability to monitor the effectiveness of some habits and routines, engenders selection pressures among these habits. If regularly attending office hours turns out not to yield dividends in terms of learning, a metacognitively-aware student may provisionally conclude this this habit should be pruned from her routine. Combinations of habits, understood by the student as routines, are also subject to selection pressures. Routines may be substituted wholesale for other routines as the student responds to information about effectiveness and environmental cues, such as suggestions from others. Selection pressures increase as opportunity costs - the

implicit price of not doing something else instead - rise with tightened scheduling. In other words, once the student concludes that she has a total of four hours to dedicate to mastering a new chapter of chemistry, this increases the pressure to select an optimal academic strategy.

What metrological consequences follow from the realization that the construct is complex? First, simplistic attempts to measure the construct set the measurement enterprise up for failure. If measurement is a primarily epistemic activity aimed at knowing, then simplistic attempts to measure complex constructs are a mismatch of epistemology and ontology. Perhaps the easiest way for a measure of academic habits to miss the mark is by having too few indicators. The second consequence of the complexity of a construct is that multiple models of the same construct will likely be required to understand all its operations. All models, even those with many indicators, involve considerable simplifications, which usually come in the form of statistical and mathematical assumptions. When undertaken deliberately, such simplifications are part of the reason for using models in the first place (Box, 1976). However, complex constructs are unlikely to be fully understandable through the lens of only one of the possible models. This insight motivates the recent trend of reporting model comparisons for complex systems, such as epidemics (Den Boon et al., 2020). In the case of academic habits, employing the same model in every attempt to investigate the phenomenon risks routinely overlooking key operations of the construct. While this study represents one attempt to model academic habits, it follows from the complexity of this construct that future attempts to design alternate models will be desirable. It seems unlikely that a single decisive model of academic habits will be arrived at, and I argue that the complexity of the construct raises the question of whether we would want such

a model. In addition to the model employed in this study, network models and agent-based models may reveal new operations of the construct.

3.4 The Qualifier "Academic"

The definition of the term "academic" used in this study refers to behaviors undertaken for the proximal goals of success in college. In higher education settings, the most commonly used measures of academic behavior are derived from Vince Tinto's Theory of Student Departure (Braxton & Hirschy, 2005), which aims at explaining higher education completion and dropout. This theory spawned over 700 studies of persistence, creating what has been called the Tintonian Dynasty (Bensimon, 2007). Tintonians argue that the decision to persist in school is driven by two major school factors: social and academic integration. Both of these constructs are commonly defined in terms of subjective belonging, that is, normative fit between students' own values and those of the institution, in both social and academic domains. Tintonians point out that, in studies of persistence, both forms of integration are consistently shown to matter, and in some cases, higher levels of one form of integration can compensate for lower forms of the other (Pascarella & Terenzini, 1983; Stage, 1989).

A number of challenges to the Tintonian model have arisen in the last four decades. Within a strictly quantitative frame, some critiques concern the model's failure to replicate in some settings and for some populations, such as students in community college (Vorhees, 1987), prompting Tintonians to respond that background characteristics are more important for persistence for these populations (Tinto, 1993). However, the larger the study, the more likely that academic and social integration appear to influence persistence (Wortman & Napoli, 1996). Perhaps a far more important issue remains that, in the words of Regina Deil-

Amen, "most attempts to validate Tinto's model more generally… do not specifically address the validity of social and academic integration as valuable concepts" (2011, p.56). The value of social and academic integration has been debated by those who argue that the Tintonian model presupposes that a disconnection from community must occur before school integration can take place, a perspective that is particularly injurious to minority students (Guiffrida, 2006).

While a complete discussion of Tintoism and its detractors is beyond the scope of this paper, there is obviously a great deal to learn from this scholarly dialogue. One source of difficulties with the Tintonian model appears to be that its central constructs - social and academic *integration* - are psychological attributes that are not meant to exert direct influence on attainment (Deil-Amen, 2011). Defining a latent psychological construct entirely in terms of behavior, as is often the case in Tintonian models (Hurtado & Carter, 1997) is arguably an example of residual operationalism in educational measurement.[20] Finally, theorists have questioned the extent to which social and academic integration are orthogonal categories, with some researchers arguing that, in practice, areas such as peer interaction about academic matters (Cole, 2007) are just as crucial. In extensive qualitative research on persistence factors among 2-year community college students, Deil-Amen found that "Socio-academic integrative moments were cited most frequently by students across all 14 two-year colleges as precursors to their persistence" (2011, p.82). With these definitional challenges in mind, it is worth revisiting the constructs of social and academic behavior

[20] There seems nothing controversial, however, in defining a latent behavioral construct in terms of reported behavior, particularly at the population level. For example, it makes sense to postulate the latent variable "spending behavior" and then sample a small domain of indicators such as "how many times a week do you eat at restaurants?". This saves the trouble of carefully reviewing all economic transactions that participants have made in the last several months.

among undergraduates using discovery-based methods. Arguably, a lower level of “pre-specification” (Wilson, 2004, p.50) or a higher level of "abstraction" (Schensul & LeCompte, 2012) – that is, fixing fewer methodological decisions in advance – would focus on the accurate measurement of behaviors before making inferences about psychological constructs.

3.5 Ontology of the Construct

Construct definitions help identify constructs among a variety of other happenings and distinguish them from other similar constructs. This form of definition, while invaluable, can fall short of handling thornier philosophical issues that arise in the investigation of a construct. Definitions can be bootstrapped from terms with a consensus interpretation, such as "academic", and thereby bypass discussion of the ontology of these terms. To allow definitions to be built up entirely within the safe confines of consensus reality may be quite alright for daily experience, but this falls short of the rigor required for inductive investigation. The definition of a "thermometer" as "an instrument for measuring and indicating temperature" exists within the confines of consensus reality, but this definition would be unsatisfactory for a group of people attempting to discern the necessary and sufficient conditions for thermometers to function. I argue that, at the very least, educational researchers should seek to articulate the necessary and sufficient material, social, and volitional conditions for focal constructs. Where we fail to do this, the locus and causes of constructs remain open to misinterpretation and misuse.

The necessary material conditions for academic habit complexity include persons, a functioning academic program, curricular and personal supplies, and time allotted for the completion of educational activities. Absent any of these material necessities, it would be

inappropriate to seek evidence for the construct of habit complexity. The functioning academic program need not necessarily be a conventional school setting, but a dysfunctional academic program can render the construct of academic habit complexity inapplicable. Sophisticated study behaviors are only rational and productive in interaction with an academic environment. Likewise, students lacking curricular supplies and time for schoolwork cannot develop academic habit complexity. Academic learning contains an irreducible material dimension, requiring the assembly of a variety of implements and tending towards the partial transformation of spaces into learning environments. Temporal extension is a necessary condition for engaging in any behavior. The building up of habits comes with additional temporal demands, since it requires gradual modifications of patterns of activity.

The necessary social conditions for academic habit complexity include a functioning academic program (again) and the social construction of academic habits as a goal. A functioning academic program is named as both a material and social necessity for academic habit complexity. The social dimension of a functioning academic program supervenes on its material dimension, and it is quite conceivable that some materially "functional" academic programs will lack the social requirements to develop of their habit complexity. Functional academic programs are thus examples of linguistic, social, and substantive (LCS) patterns (Mislevy, 2018). One of the unavoidable social features of a functioning academic program seems to be the social construction of academic habits as a goal. These may be habits of reading, of exploration of the natural and social world, of emotional awareness, of practicing a skill, and so forth. This social construction of academic habits is accomplished by drawing attention to the behavior of exemplary individuals, by direct instruction, by loops of assigned

tasks and evaluations, as well as other, subtler means. Only when this social construction has occurred, can persons appropriate the formation of these goals for themselves. Once the goal of building academic habits has been taken up by students, continued social support for academic habits is optimal (although not logically necessary) to maintain the dynamic interaction between the individual and the curriculum. This continued interaction with the student is helpful to optimally maintain the social meaning and coherence of these habits.[21]

In addition to these material and social conditions, volition is also necessary. Since academic habit complexity is characterized by purposive repeated actions, persons choose to complete each individual act and choose to complete these acts in a way that is helpful for achieving their goal. Persons choose to assent to the overall academic goal as well. In other words, many small choices are nested inside of the greater choice to pursue learning. Absent a commitment to this larger goal, choices about individual academic behaviors exist outside a motivational structure. The volitional nature of academic habit complexity explains part of the variation between individuals on the construct. Another source of variation in the construct is non-volitional, pertaining to the affordances available to the student within a social structure - for example, despite attempts to do so, not all students appear able to find a study group. Volition remains a necessary but insufficient condition for academic habit complexity precisely because willing is not enough - material and social conditions must also necessarily obtain. In the account given here, academic habit complexity has many necessary conditions and lacks sufficient conditions. This asymmetry between necessary and sufficient conditions is typical of most social constructions.

[21] None of the discussion of the necessary social conditions for the development of academic habits is meant to exclude the possibility of autodidaxy. Academic habits can certainly be developed independently. The "academic program" may be a textbook of math problems with answers in the back, and the "social construction of academic habits as a goal" may actually be accomplished via the written word.

3.6 Construct Map

As part of the construct definition phase, a construct map was created to depict the hierarchical relationship between different levels of academic habit complexity, informed by the findings of the ethnographic observation and interviews detailed in the previous phase. Figures 3.1 shows the primary construct map, which compares person characteristics with item characteristics (Wilson, 2004).

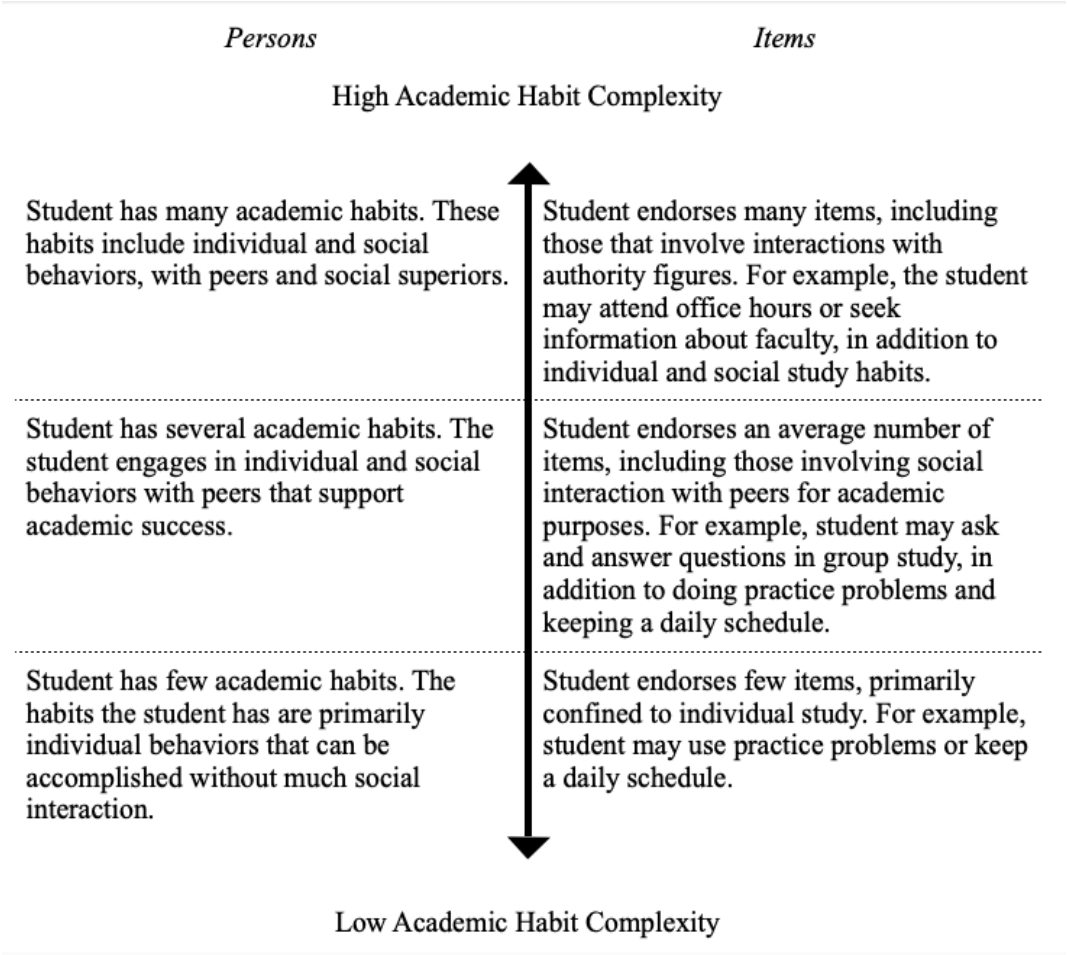


Figure 3.1: Construct map for academic habit complexity.

This map clarifies the kind of variation expected in habit complexity and among individuals who demonstrate differing levels of the behavioral construct. Later, the construct map is used to order observation statements into a continuum of item difficulty in the Evidentiary Item Map, as will be shown in the subsequent chapter. To this extent, the construct map shown here is used differently than the classic construct map, which serves as a direct guide for the generation of test items. In theory, any participant in any context that meets the necessary conditions laid out in the Ontology of the Construct can be placed on the construct map of academic habit complexity. Likewise, any item that is an indicator of Academic Habit Complexity can be placed on the map as well.

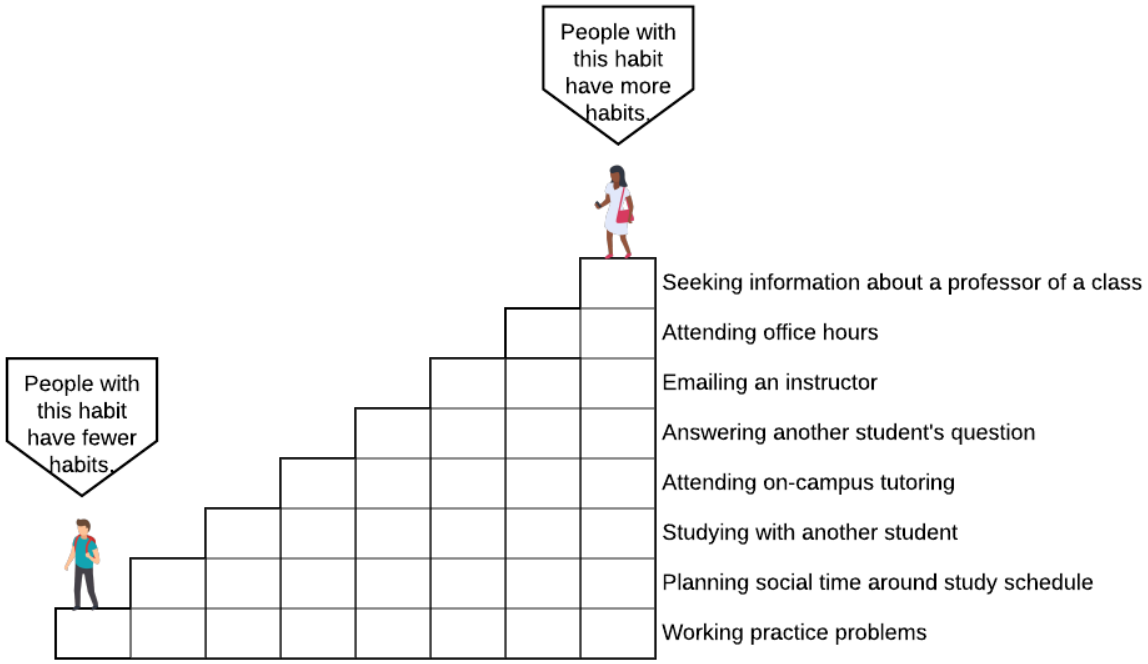


Figure 3.2: Persons and Habits. Representation of the conceptual relation between persons, number of habits, and "difficulty" of habits.

The construct map embodies several hypotheses about the relation between persons and items. First, some students have more complex habits than others. Second, some habits are harder to engage in than other habits. Third, students with more habits are more likely to engage in the harder habits and roughly equally likely to engage in the easy ones (Figure 3.2). The habits that are expected to be easiest are the ones requiring only individual behavior. Studying with practice problems is intellectually demanding, but is easy from a psychosocial perspective. The habits that are expected to be the hardest are those that require interaction with social superiors, such as instructors. In between, the habits expected to be of medium difficulty include academic interactions with peers.

3.7 Measurability

A critical issue in the philosophy of measurement is to distinguish between attributes that are measurable and those which are not. While advances in psychometrics in the 20th century gave hope to researchers who seek to measure human traits, states, and behaviors, these methodological innovations also served to sharpen philosophical issues pertaining to measurability. There are at least three major approaches to the issue of measurability that each suggest their own standards: Stevens' view, the classical theory of measurement, and latent variable theory. S.S. Stevens argued that measurement is simply the "assignment of numerals to objects or events according to rules" (Stevens, 1946). This permissive account of measurability implies that nearly any attribute can be measured. The classical theory of measurement (Michell, 1997), by contrast, reflects thinking about measurement that dates from antiquity and is shared by many today in the physical sciences. In order to be measurable, attributes must necessarily allow for the estimation of "ratios of some

magnitude" (Michell, 1997). Attributes that are not "quantitatively structured" - possessed of or reducible to characteristics such as additivity (Michell & Ernst, 1997) - cannot be measured. Latent variable theory (Markus & Borsboom, 2013) asserts that attributes cannot be directly measured (in the way that length is measured by a ruler) and that there is a probabilistic function that describes the relationship between observed indicators and the unobserved distribution of the latent variable. In order for attributes to be measurable, their indicators need to fit a hypothesized model after being translated by the link function and accounting for error. The relation between indicators and the attribute is treated as a defeasible inference that requires backing in the form of validity evidence - a lack of validity evidence undermines claims of measurability.

These three approaches to measurement suggest distinct high, medium, and low standards for the measurability of attributes. Being wary of a jump to the middle solution (i.e. the Goldilocks fallacy), there are still good reasons to select latent variable theory as an optimal standard for measurability. Steven's (1947) standard of measurability is widely dismissed by psychometricians as too permissive, since it does not require any characteristics of attributes. It is much more difficult to challenge the classical theory of measurement, as Michell's rejoinders to vigorous objections have demonstrated. Michell has argued that even robust measurement practices such as those of Rasch measurement theory and experimental manipulations of stimuli (Michell, 2008) fall short of the necessary conditions of measurement. For the present purpose, I argue that the formal structure of Michell's procedure for measurement is incompatible with a discovery-based workflow: the procedures that will ultimately be used to assess the measurability of an attribute are only appropriate to implement once a trustworthy means of gathering data about the attribute has been

established. There is a dialectical relation between the ontology of the attribute (its quantitative structure) and our epistemic means of coming to know this ontology (measurement techniques) that must be allowed to play out before the measurability of the attribute can be fairly adjudicated. Latent variable theory, with its emphasis on accumulating validity evidence and its epistemic humility that treats the attribute as unobservable, offers the right combination of checks and flexibility for this dialectic to unfold. As Mislevy (2018) has argued, latent variable theory can also be strengthened by the consideration of sociocognitive factors that shape the response process. Rasch measurement theory also strengthens latent variable theory by formalizing high measurement standards into its own characteristic workflows.

Nonetheless, Michell's overall line of inquiry remains a worthy one, namely: what characteristics of the focal attribute indicate that the attribute really might be measurable? I suggest that the avenue in which to specify these characteristics is in the construct definition phase of the measurement workflow.

The attribute of academic habit complexity has several characteristics that give cause for optimism about its measurability. First, the domain of academic habits commonly occurring in the population is likely finite. There are many, but not innumerable academic habits. Second, academic habits require time and must be completed in a timely fashion to have the desired effect. Time is quantitatively structured in the strong sense required by fundamental measurement (i.e. time is additive). Third, the pursuit of academic habits leaves behind distinct memories of events that are countable. Students have little trouble remembering whether, in the recent past, they have attended office hours or spent time completing practice problems. These three characteristics of the attribute suggest that it is more straightforwardly

rendered in mathematical terms than other commonly-measured latent variables, such as "life satisfaction" or "reading ability." The elements composing academic habit complexity are in principle reducible to a countable domain of events occurring within a defined time frame. In this sense, academic habit complexity is no more difficult to measure than the number of three-point shots made by a basketball player in two minutes.

There are, of course, characteristics of academic habit complexity that present challenges to the claim of measurability. Chief among these is the substantive "complexity" which itself serves as a warning that model-based simplifications may be unwise. It is at this juncture that Mislevy's (2018) recent theoretical work offers a crucial update to the philosophy of measurement. To those who would argue that psychometric models are simply too simplistic to show us anything about complex systems, Mislevy responds:

> [Y]es, latent-variable models, including IRT, originated under the belief that psychological variables exist in much the same way as length and force do. However, the symbol-system structures within these models can be used to express certain regularities that emerge from complex systems, across certain times, places, persons, and social interactions. They are limited in what they can tell us about the conditions under which those patterns arose, but given those conditions, they can guide reasoning locally. (pp.334-335)

While it is true that students' overall use of academic habits is a rugged landscape with many feedback loops, balancing mechanisms, and thresholds, this complex system also behaves in some reassuringly patterned ways. Indeed, such systemic patterns mark the difference between random events and complex systems. Measurement of complex systems in the classical sense favored by Michell is probably impossible. Mislevy suggests that, for complex systems, *approximation* of measurement is accomplished by "as if" reasoning in a latent variable framework (2018, p. 136). Critically, I argue that the word "complexity" is not a blank check to engage in reckless latent variable modeling. This is because inferences from

observable indicators to the unobserved attribute will require all the more backing. It is this backing that a discovery-based workflow seeks to offer at the timeliest methodological phases.

3.8 Normative Argument for Inquiry

As indicated in the previous chapter on the Construct Selection phase of measure development, normative concerns are an unavoidable aspect of educational research. Psychometricians have largely accepted that the consequences of instrument use constitute a moral issue (Messick, 1989), even if they remain divided about what this moral issue means for the validity of measures (Borsboom et al., 2004). The argument-based approach to validity which gained popularity in recent years (Kane, 1992) is helpful for framing many issues, but says little about normativity in general. The practice of making a positive normative case for the use of a measure during its development would seem an adequately response to the moral issues by educational measurement. In the case of academic habit complexity, I argue that both students and institutions benefit from the development of a measure.

Habits and routines are a salutary adaptation to the external pressures of school life, particularly during periods of transition. From the perspective of the student, school life is a ceaseless procession of new and varied happenings that require active engagement. Students are expected to form new personal relationships with instructors, to learn to appreciate new disciplines, and to generate new kinds of work products approximately every few weeks. In the eye of this storm, habits and routines are a refuge of autonomy and control. Metacognitive monitoring of one's own behavior is the only way to connect agency with

outcomes: the only kind of successes from which students can learn are the successes that follow the application of a strategy. When students decide to learn from these successes, they may decide to make strategies into habits and bundle habits into routines. In an environment full of unpredictable variation, a winning routine allows students to experience "the joy at being the cause" (Groos, 1901) - that is, the elemental pleasure, evident in infant play, at knowing that one's actions have brought about an intelligible result. Research into academic habits can inform the advice given by teachers and mentors about this common area of inquiry.

So goes the argument for studying habits. But what of the "complexity" of habits? We might, for example, attempt to locate the smallest cluster of habits that explains the most variation in some outcome, and then prescribe that students should no longer bother with the others. Why not seek out the single most effective academic habit, or some special number of habits, like three or seven? Such queries would be possible using the data gathered in this project. It is a normative decision to either count habits or seek the most parsimonious habit combination. I argue that judging overall habit complexity captures an important aspect of engagement with higher education, namely that flexibility within a designated routine is better than rigidity. Due to the changing and dynamic nature of the academic environment, students cannot afford to be like Kant: awake at 5am, two cups of tea and a smoke, lecture from 7-11am, followed by a precisely-paced afternoon walk. During an interview, one participant responded to a question about creating a study schedule: "I've tried it, and I've noticed how hard it can be to like stay on that schedule because some things can come up... like, I like having certain things in mind to have at a certain time or around a time, but if I have it like set at a time then it kind of stresses me out to like put things into it." Habits and

routines should bend but not break, as this student's experience with over-specifying the daily schedule attests. The idea of complexity allows for flexibility in crafting one's routine, but it also allows for individual latitude in determining optimal routines. As Aristotle's analogy goes, the diet of a professional wrestler will not be appropriate for the amateur athlete (Nicomachean Ethics, Book II, 1106b). Students will gain optimally from different habits - a feature of habit complexity that surpasses attempts to model the phenomenon at the aggregate level, but which is not totally disregarded by the label "habit complexity."

I argue that institutions likewise benefit from the study of academic habit complexity. The necessary social conditions for the development of academic habit complexity include a functioning academic program and the social construction of academic habits as goals. Maintenance of social support for these goals is clearly optimal. The complexity of academic habit complexity is mainly contributed by this social dimension. While autodidaxy is always possible, self-teaching outside of a community of scholars can only attain a limited complexity. Autodidacts can read, write, sketch, solve practice problems, and build prototypes, but they cannot engage in regular conversation with scholars who are either more advanced or who are at the same stage of learning (Illich, 1971) - to do this would end their isolation by definition. Consequently, academic habit complexity emerges as one of the affordances of well-structured educational communities. It is part of the value added by in-person experiences at schools and universities. As such, the complexity of our academic habits would seem to be within the classic wheelhouse of educational research, alongside effective teaching practices and educational policy (Pring, 2000).

References

Aristotle. (1985). *Nicomachaen ethics* (T. Irwin, Trans.). Indianapolis, IN: Hackett.

BiPM, I. E. C., IFCC, I., IUPAC, I., & ISO, O. (2012). *The international vocabulary of metrology—basic and general concepts and associated terms* (VIM). JCGM, 200, 2012.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, 111(4), 1061–1071.

Bridgman, P. W. (1927). *The logic of modern physics*. New York: Macmillan.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791-799.

Cobb-Clark, D. A., & Schurer, S. (2012). The stability of big-five personality traits. *Economics Letters*, 115(1), 11-15.

Cuevas, J. (2015). Is learning styles-based instruction effective? A comprehensive analysis of recent research on learning styles. *Theory and Research in Education*, 13(3), 308-333.

Den Boon, S., Jit, M., Brisson, M., Medley, G., Beutels, P., White, R., ... & Hoogendoorn, M. (2019). Guidelines for multi-model comparisons of the impact of infectious disease interventions. *BMC medicine*, 17(1), 163.

Groos, K. (1901). *The play of man*. Trans. E.L. Baldwin. Appleton.

Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds.). (2009). *Handbook of metacognition in education*. Routledge.

Illich, I. (1971). *Deschooling society*. New York: Harper and Row.

Kripke, S. A., 1980, *Naming and Necessity*, Cambridge MA: Harvard University Press.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, 47(4), 2025-2047.

Liu, J. (2017). *Neo-Confucianism: metaphysics, mind, and morality*. John Wiley & Sons.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of mathematical psychology*, 1(1), 1-27.

MacIntyre, A. (2013). After virtue. A&C Black.

Markus, K.A., & Borsboom, D. (2013). Frontiers of test validity theory: Measurement, causation, and meaning. New York: Routledge.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, 18(2), 5-11.

McGrane, J. (2015). Stevens' forgotten crossroads: The divergent measurement traditions in the physical and psychological sciences from the mid-twentieth century. Frontiers in Psychology, 6, 431.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88, 355–383.

Michell, J. (2008). Rejoinder. *Measurement: Interdisciplinary Research and Perspectives*, 6, 125–133.

Michell, J., & Ernst, C. (1996). The axioms of quantity and the theory of measurement [Part I]. *Journal of Mathematical Psychology*, 40, 235–252.

Morin, E. (2015). *Introduction à la pensée complexe*. Le Seuil.

Pring, R. (2000). *Philosophy of Educational Research*. Continuum.

Proust, J. (2013). *The philosophy of metacognition: Mental agency and self-awareness*. OUP Oxford.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, 103, 667–680.

Wagner, J., Lüdtke, O., & Robitzsch, A. (2019). Does personality become more stable with age? Disentangling state and trait effects for the big five across the life span using local structural equation modeling. *Journal of personality and social psychology*.

Wittgenstein, L. (2001). *Philosophical Investigations: The German Text, With a Revised English Translation*. 1953. Trans. GEM Anscombe. 3rd ed. Blackwell.

Chapter 4

Instrumentation

4.1 Introduction

The content of items is often identified as a source of trouble for measurement in the social sciences. For example, items may not address the full range of the phenomenon under study or may refer to the phenomenon in words that are not typically used by participants, leading to downstream difficulties in interpreting results. A great many guidelines for the construction of self-report instruments to measure attitudes and behaviors have been proposed since the rise in popularity of questionnaires in 20th century social science. Under empirical psychometric testing, some of these classic guidelines have been shown to be important, while others have been shown to be incorrect or at best ambiguous (Goretzko et al., 2019). One hallmark of these guidelines is the extent to which they mix superficial pointers about item characteristics (e.g. syntactic recommendations, reverse coding) with more profound principles issuing from the requirements of measurement, such as clarity concerning the intended response process. I refer these latter principles as "design principles." The general purpose of such principles is to provide a conceptual bridge between the construct and the plan for the instrument, transitioning into the normative mode.

4.2 Design Principles

While this list of principles is not exhaustive, I claim that self-report survey instruments require at least these five:

1) *Reference principle*. The domain of the attribute should be coextensive with the domain of its observed manifestations. The content of items in self-report instruments

should refer to behaviors that members of the focal population have been observed doing or thoughts, attitudes, and beliefs that they report experiencing. To support inferences about these behaviors and attitudes, descriptions of participant communication and behavior considered should be detailed and interpretative.

2) *Access principle*. Items should be presented in such a way that the vast majority of participants can understand and evaluate them - that is, the form taken by the items should not be idiosyncratic. Self-report items should not include elements, including answer choices, that are difficult for participants to interpret due to construct-irrelevant factors (Messick, 1995). Instruments should be designed so that the maximum number of members of the target population can participate.

3) *Authenticity principle*. Of the many possible forms which the content of a self-report item might take, the optimal form is the one most similar to the way the participant articulates or experiences the referent(s) of the item.

4) *Investment principle*. Participation in measurement should necessitate cognitive effort commensurate with the expected benefit of participating in measurement. More demanding tasks should be associated with greater benefits, and no intrinsic benefit should be assumed. When few benefits are available, instruments should strive to limit cognitive effort as much as possible.

5) *Inferential principle*. Measurement should involve traceability chains that show how all inferences of which researchers are aware are justifiable. The content and form of items are important parts of this traceability chain. Domain analysis is the foundation of measurement, and the links between it and other layers of assessment design should be clear (Mislevy et al. 2003).

Each of these five design principles begins with an ontological characteristic of measurement and arrives at a normative statement about what the designer should do. The Reference Principle begins with the claim that there is variation within observed manifestations of constructs and arrives at the normative claim that instruments should capture as much of this observed variation as possible. Overly narrow "operationalizations" are normatively excluded by this principle. Both the Access Principle and the Authenticity Principle are justified by beginning with the ontological claims that 1) response processes are the causal chain between the construct and the instrument (Borsboom et al., 2009) and 2) mental models of the construct may be population-bound. The Access Principle proceeds from these observations to the prescription that the design of the instrument should reflect the mental models of as large a proportion of the target population as possible. This principle requires that items should be communicated in terms that are as general as possible within the population and prohibits items that some members of the population do not understand. The Authenticity Principle counterbalances this principle with the requirement that items must be closely tailored according to evidence about the mental models of the participants. This principle prohibits contestation of the mental model of the participant during the measurement process, for example, by implicitly substituting a medical model of the construct for the common model. The Investment Principle acknowledges that the response process is at base a form of mental and physical exertion and that the energy of participants is limited. It prescribes that these exertions should be carefully monitored and wherever possible, rewarded. It prohibits the use of laborious instruments in cases where the return on investment is likely to be small or nonexistent. The Inferential Principle begins with the ontological claim that measurement

instruments must be traceable in order to be evaluated and arrives at the norm that all phases of instrument development should be clearly documented without gaps, and that warrants should be collected for the appropriateness of activities at each stage of development. Black boxes are undesirable in measurement (Maul et al., 2018), and instruments that rely on conspicuous black boxes may not be justifiably called "measures."

With these design principles in hand, the instrumentation phase of the workflow gains a normative foothold. When tradeoffs must be made in instrumentation, returning to these design principles provides necessary reminders about the proximal goal of the human measurement.

4.3 Instrument Format

Having selected the construct of academic habit complexity for measurement (Chapter 2) and established a conceptual definition (Chapter 3), the aforementioned design principles were consulted to build a measure that would be generalized to the larger population of first-year pre-biology majors. Decisions about the format of the instrument can be described in terms of the design principles. The principles are used below to categorize decisions about item content. Explanations of the rationales for these decisions about instrument form and content is intended to satisfy the Inferential principle.

The Reference Principle, which holds that full construction representation should be the goal, was key to determine the scope of the measure. Given the construct definition, information about a broad variety of behaviors needed to be collected. Questionnaires are serviceable tools for cuing participants about a large number of behaviors. The format of the survey task was selected to capture participation in a diverse array of social and individual

practices. Moreover, the construct needed to be represented temporally. Self-report was selected as the optimal information source for detailed information about behavior over an extended period. The time frame selected for the task was based on ethnographic observations of behavioral change among first-year students in the program, who appeared to change their behavior over the course of more than two weeks. For example, Raúl's realization that he needed to improve his time management appeared to occur within a 3-week span, as did Mika's switch from studying alone to studying in a small group, an observation consistent with literature on habit formation, which proposes 18 days as a minimum for habit formation (Lally et al. 2010). The survey was administered near the end of the quarter, a period in which student behavior may change due to final exams. It was postulated that two weeks would be enough time to capture typical final exam preparation behavior as well as more typical behavior, a judgment supported by the ethnographic observation that few students had begun studying for final exams in the final week of the quarter, which occurred several days after the survey was administered. If the survey had sought only to capture exam-motivated behaviors, for example, by asking about just the few days just prior to the exam, the construct may not have been fully represented[22]. The items were designed to vary along a continuum of difficulty to account for the observation that some habits were apparently easy to implement while students struggled with others. Intuitively, it seemed unlikely that students would have either no academic habits or more than 22 distinct academic habits. The varying difficulty of these behaviors were thus postulated to vary according to a roughly sigmoid function, with upper and lower probabilistic limits of difficulty.

[22] Later administrations of the survey have queried participants both at the middle and end of the term, in order to track habit complexity throughout.

The Access Principle requires that participants be able to understand and evaluate the format of the instrument, while the Authenticity Principles states that the optimal format of the instrument is closely tailored to formats that are culturally congruent for participants. Checklists are in common use among individuals in target population, and are employed for purposes ranging from grocery shopping to academic work. By comparison, other formats such as semantic differentials (e.g. strongly disagree to strongly agree) may lack analogs in the daily experiences of participants. Participants were not required to understand the meanings of any levels of the same item (e.g. as in a Likert scale), a common source of trouble from the perspective of the response process. Since the items were dichotomous (yes/no), participants responded to the instrument by simply clicking on the text of the item in the checklist. The survey interface was clean and minimalist. The Access Principle also guided the selection of the online survey over a paper survey, since the online surveys made with the Qualtrics platform are compatible with accessibility tools the participant regularly uses for computer-based work, such as personalized modifications to text size, read-aloud software, braille displays, eye typers, and puff-suck switches. All students at the university have access to personal computers, public computers, or rentable computers, and the majority of students use multiple digital devices on which a short survey might be taken comfortably. Indeed, checklist instruments are much easier than ordinal scales to complete on mobile devices, since checklists wrap more smoothly in web pages that are responsive to screen size. As Peterson and colleagues (2017) remark in their review of smartphone survey practices, in all the "excitement" about using smartphones for data collection, "it seems that researchers never imagined that their long and complex surveys would be completed using pocket-sized devices" (p. 203). Using a smartphone to participate in survey research does not appear to

lower data quality, but *in vivo* response process data reveals that complicated survey interfaces are a source of trouble for mobile users (Antoun et al. 2017). As a large proportion of surveys are now taken on small devices, designing instruments with these interfaces in mind is increasingly important for both Access and Authenticity.

Since the marginal benefit of participating in the study was small, the Investment Principle dictates that the amount of effort exerted by participants should be minimized. Several features of the instrument were determined based on this principle. First, a simple checklist, dichotomously scored, was chosen as an appropriate and participant-friendly format. To complete the exercise participants tapped or clicked a digital button on which the academic habit was presented and clicked once to turn the page. The items were broken into two separate pages to make scrolling easier. No rankings (e.g. Likert scales) or semantic differentials were required. Second, the two-week time-span limited cognitive effort as much as possible while still honoring the temporally-extended nature of the construct (see Chapter 2). Finally, participants were offered the chance to win gift cards for participating in the survey, increasing the expected marginal benefit of participating in the study. Third, identities of the participants were blinded from the perspective of the researchers, and participants were informed that participation was confidential.[23] This assurance decreased potential consequences of participation in the survey, such as being contacted by university faculty and staff about responses, lowering the overall stakes of participation.

[23] All data were matched with student grades and de-identified by the UCSB Office of Institutional Research.

4.4 Item Content

Just as the design principles act as a bridge between the characteristics of good measurement and the normative prescriptions to be followed in the structure of the instrument, the content of items can also be described as an application of these principles. Again, I argue that the only way to satisfy the Inferential principle is to fully describe all design decisions in the workflow and their relation to such normative principles.

The Reference principle was enacted via the construction of a knowledge representation (Markman, 1998; Mislevy & Riconscente, 2011) I have called the Evidentiary Item Map. The evidentiary item map includes illustrative examples, references to fieldnotes in which the observation is documented, references to course documents in which the practice is referenced, and the item text (Appendix II). This procedure is meant to render item authorship more transparent, so that critical decisions about instrument construction are justified rather than black-boxed. Uncovering and making explicit the qualitative judgments that undergird quantitative studies is an important project of mixed methods designs, and tools like the Evidentiary Item Map are suggested as part of a rigorous multi-method research design (Turner et al., 2017).

Since all items refer to utterances and actions taken by participants and recorded in field data, the Evidentiary Item Map is also meant to satisfy the Authenticity principle. Actual examples of participant activity were consulted during the authorship of each item. The local, emic lexicon (Pike, 1967) was used whenever possible. For example, specific campus services such as group tutoring sessions were referred to by the acronyms most commonly-used among undergraduates.

The Access principle was satisfied for item content by ensuring that items were of the appropriate score and that members of the target population did not face any problems in interpreting the items. While the ethnographic observations that led to the identification of a particular academic habit were usually highly specific, items were deliberately phrased in such a way that multiple observed manifestations might be reasonably categorized as an instance of this habit. Cognitive interviews (n = 27) were conducted with target population to appraise the extent to which items were understood as intended. Analysis of these interviews revealed that no participants exhibited difficulty in the comprehension phase of the response process (Tourangeau et al., 2000). Participants were able to generate utterances similar to item text featured in the instrument, paraphrase the meaning of items, generate additional items that might be added to the instrument, and sort all items into piles based on difficulty. The subsequent chapter explores findings from the cognitive interviews in greater detail.

The Investment principle suggests that items should be authored to make the amount of effort required to respond to them commensurate with the marginal benefit of doing so. If items are written in such a way that the response process induces high effort or discomfort, the expected benefit to participant should be greater. In situations in which few marginal benefits are likely, items should require minimal effort on the part of the participant. This latter scenario was the case for the AHCS, and items were authored to minimize strain and discomfort. No items were included that might require participants to disclose sensitive information. One item queried whether participants had taken advantage of campus-based mental health services. To reduce potential discomfort in responding to this item, mental health services were rhetorically bundled in the text of the item alongside non-mental health services - the food bank and the medical clinic. Participants responding affirmatively to this

item might theoretically have taken advantage of any combination of these three campus services. By ignoring traditional wisdom regarding double-barreled item, which has recently been shown to be less severe of a problem than previously thought (Goretzko et al., 2019), this item was written to facilitate plausible deniability of the use of a stigmatized service.

Finally, the use of the Evidentiary Item Map to track the development of item content is an application of the Inferential Principle. Since the evidentiary item map provides links to the observations that motivated the authorship of each item, a path can be traced from the results of statistical analyses back to observations, which can be used as evidence for interpretation. Explicit recursive linking serves as much-needed validity evidence for the construction of measures, specifically evidence for the validity of instrument content (AERA, APA, NCME, 2014).

To illustrate the process the process of evidentiary item mapping, it is useful to proceed through an example of an item developed in this way. This proceeded in four stages: 1) re-analysis of field notes and memos for instances of socioacademic practices, 2) description of these practices in more general propositions of appropriate scope, 3) syntactic restructuring into an item format, 4) cognitive interviewing. In the first step, field notes and memos were re-analyzed for manifestations of socioacademic practices. One practice that appeared multiple times in this analysis was the idea that students should seek information about professors before taking their class, weighing this information seriously in their decision. Some contexts in which this phenomenon occurred included the following:

> Context A. The instructor advises students "never to walk into a class on the first day without knowing how that class operates" explaining that other people are a valuable source of information. (Week 1)

> Context B. Raúl explains that he has realized that students who get lower grades tend to review professors poorly. Sonya agrees and says as a result, she no longer trusts ratemyprofessor.com as a source of information. "Take what people say about a class with a grain of salt," Steve affirms. (Week 3)
>
> Context C. "I'm scared about biostats," a student says, referring to a future required class. Steve replies with lots of advice for how to plan one's course schedule. Later, he warns first-years to watch out for a pileup of difficult courses in the second year, telling them that things even out after that. He even suggests taking physics during study abroad to skirt the professors in the local physics department. (Week 3)
>
> Context D. When asked about his personal study strategies, Steve says to the whole class "Find out about the professor and what's going to be important to them." (Week 6)

From these descriptions, we learn several key bits of information about the practice of seeking out information about future classes which will aid us in Step 2, the general formulation of a proposition about these observations. We can see here that professor characteristics are considered a make-or-break factor in choosing whether to take a class, and even cause some students to time a strategic exit from the country via study abroad to avoid some faculty. Students gather information about professors from multiple sources including word of mouth and online websites, and critically weigh the quality of such information. Further, the information gathered here is considered helpful not only for course selection but for success during the class. These considerations are generative for writing an item that includes important identifying details but excludes disqualifying elements that might have been included otherwise. In this case, the item should focus on the "professor" and not just the class, should be neutral regarding the medium by which information is communicated, and should not limit the purpose of the information to course selection. Thus, in Step 3, the following item was written:

> "In the past two weeks, I have: Sought detailed information about a professor of a class I want to take"

This item was included during the cognitive interviewing procedures with undergraduate participants (n=27) and the item was found to function as expected. However, while my first draft of the item initially included the word "detailed" before "information", a cognitive interview participant suggested that it introduced unnecessary ambiguity. Once again consulting field notes and memos, it was determined that the notion of seeking information of great "detail" was not a critical part of the targeted practice, and the word "detailed" was removed, resulting in the following item text:

> "In the past two weeks, I have: Sought information about a professor of a class I want to take."

Using this four step process of re-analysis of field notes and memos, propositional description, formulation into items, and cognitive interviewing, a total of 22 items were formulated for the with 12 referring to the social academic behaviors and 10 items referring to individual academic behaviors. Once the evidentiary item map was completed, the construct map was updated with specific items. This allowed for the connection of hypotheses about item content to item difficulty. Items were expected to be easiest to endorse if they required only individual behavior (e.g. completing practice problems), of medium difficulty if they required social behavior involving peers (e.g. group study), and of high difficulty if they required social behavior involving authority figures (e.g. attending office hours or associating with academic societies, which are primarily composed of upperclassmen).

4.5 Administration

First-year undergraduates enrolled in a challenging introductory chemistry course were invited to participate in the study. This invitation was issued via email and in-person solicitation from faculty. Participants were informed that they would be entered into a drawing to win a gift card. Students who participated in the BIOME program received an additional solicitation during class. This was done to ensure that a sufficient number of participants in the targeted program would enroll in the study for a between group comparison to be possible.

Students were informed that participation in the study was confidential, and that neither participation nor non-participation would affect their grade in the course. As invitations to the study were issued in a lecture hall and by email, students had the option to complete the instrument in the lecture hall or in the privacy of their residence. The AHCS was included as part of a short questionnaire with other, similar instruments. The median time to complete the instrument was roughly three minutes.

Ultimately, 310 undergraduates elected to participate in the study. Of these 55 were participants in the target program, while the remaining students were not. The demographic characteristics of participants in the program are summarized in Table 4.1. These demographics are approximately representative of students in the biology major at the university.[24] Since one of the objectives of the BIOME program was to provide support for underrepresented minority students (URMs) in the biology major at the university, adequate representation of URMs in research about the program was considered a high priority. Every

[24] The biology program at the university enrolls 62% women (compared to 69% in this study), and 24% underrepresented minority students (compared to 28% in this study). That is, women and minority students are slightly oversampled in the current study.

member of the target population is assumed to have been equally likely to be selected for participation in the study. Three pieces of evidence serve as backing for this assumption. First, multiple strategies (email and in-person solicitation) were used to recruit students into the study. Second, these strategies were designed to reach the entire target population, including, for example, students who were absent on the day when participation was solicited during lecture. Third, the sample was demographically representative of the first-year student body at the university. These sampling strategies are helpful to avoid the necessity of weighting cases to compensate for bias in the response rate.

Table 4.1

Number of Participants in the Program and Comparison Group

	Treatment	Comparison Group
Gender	42 F, 13 M, 0 Other	172 F, 83 M, 0 Other
First-generation	30	143
Pell-eligible	20	86
Asian	13	103
URM	20	69
White	21	77

Disclosure avoidance refers to strategies to de-identify confidential data. A robust disclosure avoidance strategy was employed for this study. Students provided their university identification number on the survey, rather than their name. Surveys were processed by Institutional Research at the university and ID numbers were used to match students to their records, such as entering SAT scores and grades in the chemistry course selected as a dependent variable for the study. Institutional Research then removed student ID numbers and forwarded data to research team. Thus, student identities in the quantitative portion of the

study were fully blinded from the perspective of the researchers. The risk of "inadvertent direct disclosure" - that is, accidental reidentification of participants in the study - was reduced by omitting items on the survey of institutional data that might allow triangulation of participant identities, such as date or place of birth (Biemer and Lyberg, 2003). To guarantee that students cannot be re-identified, students who identify as Black, Latinx, Native American, Pakistani, and Filipino students have been grouped together using the label URM.

Data processing followed the recommendations of Biemer and Lyberg (2003, pp.215-257). To reduce human error, all edits of the raw data were performed using automated steps rather than manual recoding. After each step, acceptance sampling (Dodge & Romig, 1944) was performed on 3% of the data to discover any defects in editing. Acceptance sampling is an industrial quality control method in which batches of the product are removed at random and inspected individually for defects - the application of the method to cases in a statistical analysis is relatively straightforward (Biemer & Lyberg, 2003, p.221). Whenever any defects in the automation process were revealed, the dataset was returned to its previous state and the operation revised. Analysis of paradata included checks of survey completion time and missingness. Participants who took longer than 30 minutes to complete the survey, roughly ten times the median duration, were excluded as likely low-effort responders (n = 36). Likewise, participants for whom key demographic predictors (e.g. ethnicity) were missing were excluded from subsequent analyses (n = 7). Following processing, 270 cases remained in the dataset, a sufficient sample size for all subsequent analyses.

4.6 Rasch Measurement Theory

A construct map of program outcomes was hypothesized showing an ordinal progression of items by difficulty (Wilson, 2004). Once item difficulties are estimated via the Rasch model, empirical difficulties can be checked against prior expectations. For example, one hypothesis embodied in the construct map was that behaviors requiring only individual actions, such as working on practice problems, would be easier to endorse than behaviors requiring interaction with others, particularly authority figures (Figure 3.1). A goal of the model-fitting phase of measurement is to determine whether the hypotheses embodied in the construct map are sound. If they are, we should expect the model to fit, or come very close to fitting, on the first attempt.

Item-response theory, and in particular Rasch measurement theory (RMT), is used to investigate the severity and discrimination of individual items (Bond & Fox, 2015; Rasch, 1960). In RMT, a well-functioning item is one that helps to distinguish persons with a given level on the focal construct from persons with qualitatively different levels on that construct. One key difference between the Rasch model and other IRT models with more parameters (2PL, 3PL, and so on) is that the Rasch model requires that all items have the same discrimination, mathematically formalized as fixed slopes. In the semantics of the model, this means that an item is required to function equally well at detecting differences between persons with a given level of the construct, no matter the level each person. This feature of the Rasch model allows for the approximation of interval-level units (i.e. like a thermometer) rather than "units" of arbitrary length for persons with differing levels on the latent construct – an affordance which tends to match the intuitions of practitioners not familiar with psychometrics. A well-functioning scale is composed of multiple items that are appropriately

targeted to the distribution of persons on that construct. Items with varying levels of severity are desired to represent the whole relevant range of the construct, while items that most effectively distinguish persons from one another - parameterized as slope - are considered optimal. Given a scale with good psychometric properties, items that function differently across demographic groups can be flagged for additional analysis - a process which aids in the generation of theory about the system under measurement. The basic process of the fitting the Rasch model can be divided into five phases of analysis: analysis of item fit, analysis of person fit, analysis of item difficulty, analysis of person-item targeting, and analysis of differential item functioning across groups.

Item fit is the major criterion for determining whether the overall model fits (Linacre, 2002). In RMT, item fit and model fit are treated as a single gestalt: if all the items fit the model, then the model fits. Item misfit, especially underfit, is evidence that items or tasks do not fit the Rasch model and are unrelated to the hypothesized underlying continuum. RMT dictates that underfitting or overfitting items are not suitable for measurement (Bond & Fox, 2015). Underfit occurs when item responses are erratic, there are unmodeled sources of variance, or if the items do not discriminate between people with varying levels of the hypothesized construct. Overfit occurs when item responses are deterministic, and signals a surfeit of items or the presence of seriously underfiting items.

An analysis of person fit, particularly underfit, can reveal the proportion of participants whose behavior is adequately summarized by the model - that is, the behavior of persons conforms to the requirements of RMT. For example, a person with relatively few academic habits will misfit the model if she also endorses one or more items that are supposedly high in difficulty (e.g. attending office hours). When a person misfits, we know little about her

through the lens of the model. In RMT, a high proportion of misfitting persons or items is treated as evidence that the data-generation procedure or theory may be flawed. The use of person fit statistics marks one of the major disciplinary divides between item response theory and classical test theory, the users of which typically do not estimate the proportion of participants who are well-described by the model (Embretson, 1999; Magno, 2009). The meaning of person fit statistics in RMT is especially well-defined relative to person fit statistics derived from other IRT models (Meijer & Sijtsma, 2001).

In RMT, item difficulties should span the continuum of the construct. This means that items should be written to describe low, medium, and high levels of the construct. It is both a practical and mathematical problem that scales lacking in items that target a particular level of the construct have low discrimination for participants at that level. For example, a scale to measure anxiety that lacks items to measure "severe" anxiety-related behaviors will be less able to distinguish between patients with medium-high and extremely-high anxiety. This is another notable disciplinary difference between users of item response theory and classical test theory - the former take explicit steps to determine the extent to which scales gather adequate information about participants with levels that span the whole range of the construct (Embretson, 1999). In models, this span is parameterized as delta (δ), usually called "item difficulty" or "item severity."

Person-item targeting refers to the extent to which item difficulty is a suitable match for the level of the construct present in the population. Just as a lack of items targeting a given level of a construct is suboptimal for measurement, an entire scale may be "too easy" or "too hard" for the target population, and thus be unable to provide detailed information about a large number of participants. In RMT, person-item targeting is checked by comparing the

range and distribution of item difficulty estimates to the range and distribution of person ability estimates. A unique feature of the Rasch model is that person abilities and item difficulties are on a common (logit) scale. This allows for the range and distribution of item and person estimates to be visualized in a data representation known as a Wright Map (Bond & Fox, 2015). On the left side of the Wright Map, the distribution of person ability is shown as a vertical histogram, while on the right side, the distribution of item difficulty is shown via a scatter plot on the same vertical axis. When the dots on the scatterplot occur along the whole vertical range of the histogram, the instrument is well-targeted to the population. Provided the data fits the Rasch model, a well-targeted instrument with no major gaps in item difficulty is evidence of adequate construct representation, meaning that a wide range of plausible manifestations of the construct has been addressed (Messick, 1995).

Differential item functioning (DIF) refers to a scenario in which an item is easier to endorse for a person in one group than another group, despite both people having the same level on the hypothesized latent variable. When DIF amounts to more than statistical noise related to sampling or estimation (Hagquist & Andrich, 2015) it may indicate a violation of invariance in the model - analogous to a thermometer that measures the temperature of water more accurately than it measures the temperature of gasoline. In most practical applications, DIF analysis is used to flag items that may be unfair or biased (Bond & Fox, 2015). However, the model semantics of DIF are also consistent with the interpretation that items that exhibit DIF are relatively easier or harder for the focal group for reasons besides item bias (cf. Zumbo, 2007, Henson et al. 2010). This latter interpretation is perhaps more useful for modeling behaviors, which may be easier for members of some groups to access or

perform because of linguistic, cultural, and substantive patterns existing in localized settings (Mislevy, 2018).

4.7 Fitting the Model

Statistical analyses were conducted in Test Analysis Modules (TAM), an open-source R package (Robitzsch et al., 2018).[25] Mean-squared infit statistics of all items were within canonically acceptable ranges (Bond & Fox, 2015; see Table 2). To aid in the visualization of fit statistics at a glance, I take this opportunity to introduce a data representation I call a Fit Web (Figure 4.1). In the Fit Web, distances between radial strands of the web are 0.1 mean-squared (MNSQ), with the outermost strand set to 1.3 MNSQ, the canonical upper boundary for this fit statistic. The Fit Web permits the visualization of item fit at a glance - any item that does not fit will cause the polygon to spread to the edge of the web. Underfitting items cause the web to contract towards the center. The ideal model fit will spread the web around a mean-squared value of 1, three strands away from the web's edge. By comparison, models with poor fit appear as asymmetrical and jagged webs, and are easily visually identified. The introduction of this data representation (like that of the Evidentiary Item Map and Construct Representation Tree) is meant to smooth the measurement workflow and aid in communication with non-experts.

[25] All R code was written by me and all interpretations are my own. Portions of this dissertation appear in a jointly authored paper with Dan Katz submitted to AERA 2020.

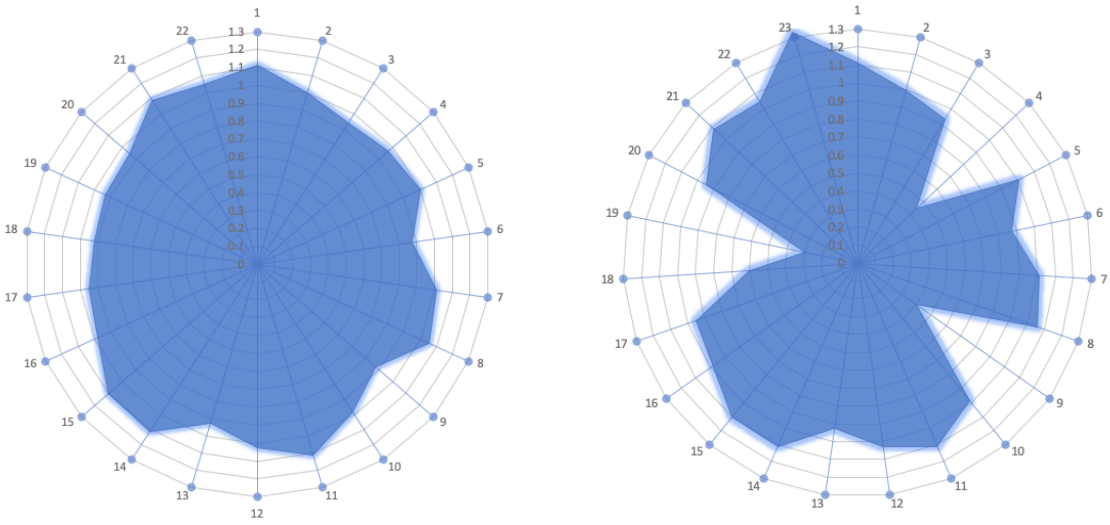


Figure 4.1: The Fit Web depicts item fit statistics as an n-sided polygon, where n = number of items. Distances between radial strands of the web are 0.1 MNSQ, with the outermost strand set to 1.3 MNSQ. This fit web permits the visualization of item fit at a glance - any item that does not fit will cause the polygon to spread to the edge of the web. The web on the left is the true Fit Web for the AHCS, while the illustrative web on the right is from a fictional scale with several underfitting items and one overfitting item.

After model fit, dimensionality was investigated. A PCA of standardized residuals of the Rasch model revealed no components with large eigenvalues, such that potential sub-structures accounted for no more than 7.7% of the unexplained variance (Linacre, 1998). Additionally, a multidimensional model was specified experimentally, and the unidimensional model was found to have superior fit to a multidimensional model using the Bayesian Information Criterion (BIC). These procedures provide evidence that the AHCS is a measure of a single attribute with minimal construct irrelevant variance.

Once item difficulties are estimated via the Rasch model, empirical difficulties can be checked against prior expectations recorded in the construct map. Item difficulties conformed to hypotheses and covered a broad range of student abilities (Figure 6). Given the theory that activities pursued individually would be easier for students, it is not surprising to see that working on practice problems was very easy to endorse, studying with others had medium difficulty, and going to office hours was among the harder items to endorse (see Table 2 for item difficulties). However, not all items conformed precisely to the hypothesized difficulties. For example, creating and following a study schedule was in the middle range of difficulty even though it is an individualized action. This suggests that there may be unmodeled determinants of item difficulty (such as the degree of planning required for an action) or that greater emphasis may need to be placed on these program goals. In addition, model-generated item characteristic curves were compared with empirical patterns of item response for all 22 items and no unusual behavior was discovered.

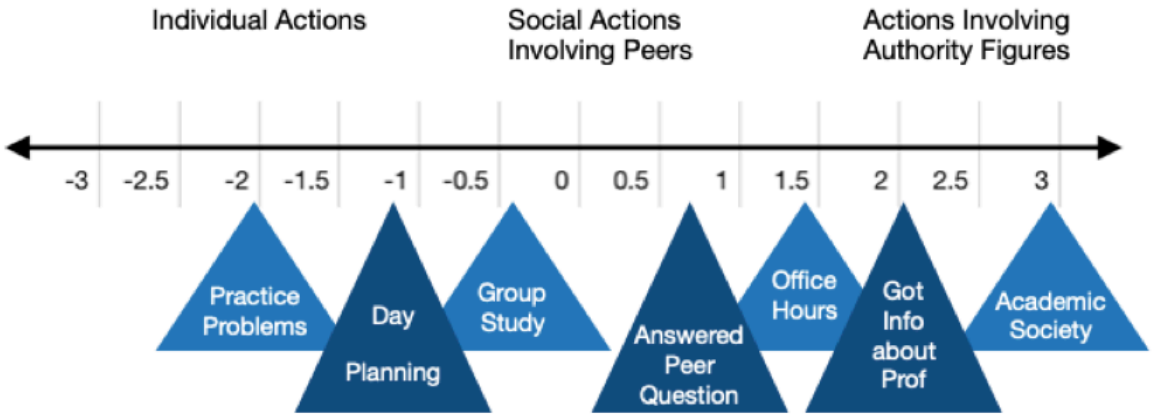


Figure 4.2: Some empirical item difficulties placed on a horizontal construct map. Doing practice problems was relatively common, while seeking information about faculty was relatively rare. Units are in logits, approximating interval-level measurement.

An analysis of person fit reveals that approximately 9% of participants underfit the hypothesized model (MNSQ > 1.3). This indicates that there is more randomness the in behavior of some persons than anticipated. A check of misfitting persons showed that participants were approximately equally likely to fit the model regardless of observed group or program membership. Groups examined were gender, underrepresented minority status, and Pell grant eligibility - a binary proxy for SES. This suggests that the model of academic habit complexity suits the population of participants, with special attention paid to their heterogeneous backgrounds. Person-separation reliability was .77, which can be interpreted in a manner comparable to the critical values of Cronbach's alpha (Bond & Fox, 2015), indicating that scores on the instrument can adequately differentiate multiple, qualitatively distinct groups of participants.

As depicted on the Wright Map, person-item targeting was generally successful. Item difficulties stretch almost the entire span of person ability, providing a high level of discrimination even for students with the most complex academic habits. The difficulty range of items was wide enough to capture a wide range of person ability, and in fact the spread of item difficulty (SD = 1.21 logits) was slightly larger than the spread of person ability (SD = .98 logits). However, person-item targeting fell short when it came to participants with low habit complexity. Fortunately, analysis of raw response patterns reveals that no participant had such a low level of habit complexity that they did not answer at least one item positively. Mistargeting of the scale was not severe: the mean item difficulty (.26 logits) was only .21 logits higher than the mean person ability (.05 logits). Future iterations of the AHCS will pilot additional items to gain greater information about students of low-level of ability.

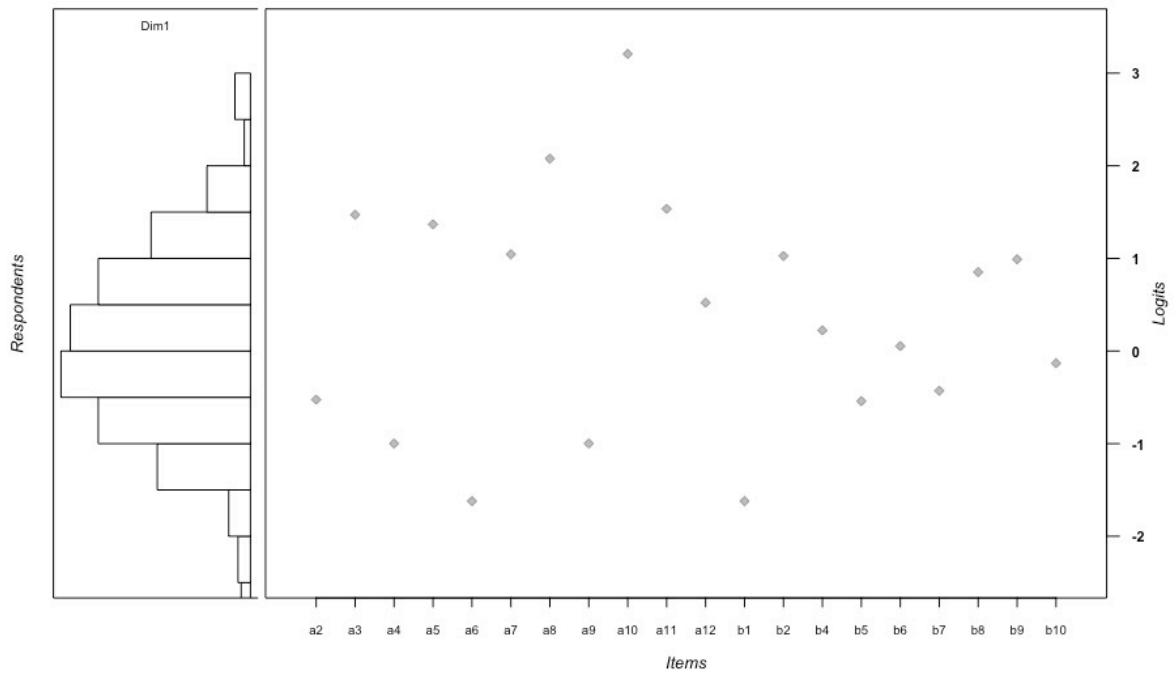


Figure 4.3: Wright Map produced by Test Analysis Modules (TAM). The map shows that the items (the dots on the right side) are well-targeted along range of the construct for participants (the histogram on the left side). Both person ability and item difficulty on a common logit scale.

Items were checked for DIF for the observed groupings of gender and URM status. Results are displayed in Table 2, below. Useful results from our analysis of DIF include the finding that women in our sample found it relatively easier than men to mark problems to return to later while studying while men found it relatively easier to plan ahead for breaks (see Table 2). These findings offer clues to the structure of student behavioral patterns and key insights for continuing development of the program. For example, marking problems and returning to them later is a strategy that may need to be emphasized for men, who were relatively .89 logits less likely to do this than women.

Based on these results, it may be reasonably concluded that the AHCS fit the hypothesized psychometric model. No items were deleted, and no changes were recommended to the content of existing items. Future iterations of the AHCS will pilot items of lower difficulty to improve person-item targeting.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*.

Antoun, C., Couper, M. P., & Conrad, F. G. (2017). Effects of mobile versus PC web on survey response quality: A crossover experiment in a probability web panel. *Public Opinion Quarterly*, 81(S1), 280-306.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. John Wiley & Sons.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Routledge.

Borsboom, D., Cramer, A.O.J., Kievit, R.A., Z, Scholten, A. & Franic, S. (2009). The end of construct validity. In R.W. Lissitz (Ed.), The concept of validity: Revisions, new directions and applications (pp. 135-170). Charlotte, NC: Information Age Publishing.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design approach*. Hoboken, NJ: John Wiley.

Dodge, H. F., & Romig, H. G. (1944). *Sampling Inspection Tables*, Wiley, New York.

Goretzko, D., Pargent, F., Sust, L.N.N., & Bühner, M. (2019). Not very powerful: The influence of negations and vague quantifiers on the psychometric properties of questionnaires. *European Journal of Psychological Assessment*, Advance Online Publication, p.1-11.

Hagquist, C., & Andrich, D. (2015). Determinants of artificial DIF-a study based on simulated polytomous data. *Psychological Test and Assessment Modeling* (Vol. 57).

Henson, S., Blandon, J., & Cranfield, J. (2010). Difficulty of healthy eating: A Rasch model approach. Social Science & Medicine, 70(10), 1574-1580.

Lally, P., Van Jaarsveld, C. H., Potts, H. W., & Wardle, J. (2010). How are habits formed: Modelling habit formation in the real world. *European journal of social psychology*, 40(6), 998-1009.

Linacre, J. M. (1998). Detecting multidimensionality: which residual data-type works best?. *Journal of outcome measurement*, 2, 266-283.

Linacre, J. M. (2002). What do infit and outfit, mean-squared and standardized mean? *Rasch Measurement Transactions*, 16, 878.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response

theory using derived test data. *International Journal of Educational and Psychological Assessment*, 1(1), 1-11.
Markman, A. B. (1998). *Knowledge representation*. Mahwah, NJ: Erlbaum.
Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied psychological measurement*, 25(2), 107-135.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9), 741.
Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.
Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. ETS Research Report Series, 2003(1), i-29.
Mislevy, R. J., & Riconscente, M. M. (2011). Evidence-centered assessment design. In *Handbook of test development* (pp. 75-104). Routledge.
Peterson, G., Griffin, J., LaFrance, J., & Li, J. (2017). Smartphone participation in web surveys. Total survey error in practice, 203-233.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.
Robitzsch, A., Kiefer, T., & Wu, M. (2018). TAM: Test analysis modules. *R package version, 2-0*.
Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.
Turner, S. F., Cardinal, L. B., & Burton, R. M. (2017). Research design for mixed methods: A triangulation-based framework and roadmap. *Organizational Research Methods*, 20(2), 243-267.
Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Routledge.
Zumbo, B. (2007). Three generations of DIF analysis: Considering where it has been, where it is now and where it is going. *Language Assessment Quarterly*, 42(2), 223–333.

Chapter 5

Validation

5.1 Introduction

Validation is a process of determining the quality of a measurement instrument by collecting evidence about its real-world performance. In education and psychology, the areas of performance about which evidence is required have been debated for decades (e.g., Loevinger, 1957; Borsboom et al., 2004, Newton & Baird, 2016). The current *Standards* for Educational and Psychological Measurement points to five areas for which validity evidence should be sought: instrument content, relationships to other constructs (criterion), the response process of participants, the internal structure of the instrument, and the consequences of using the instrument (AERA, APA, NCME, 2014).

Evidence collected about instrument content typically includes expert review of item content, focus group data from the target population, or reference to curricular standards. Criterion validity evidence is typically gathered using regression to determine the extent to which a measure correlates with scores on a desired outcome or another measure of the same construct. Validity evidence about the response process is typically collected from cognitive interviews or user behavior while taking the instrument, such as eye-tracking and mouse movements. Validity evidence about the internal structure of the instrument refers to an analysis of the output of the psychometric model, such as item-level statistics, reliability coefficients, and analyses of dimensionality. Validity evidence of the consequences of instrument use is the least defined category from a disciplinary perspective, but sometimes includes evidence of fairness, classification accuracy, or follow-ups with participants with whom the instrument was used.

Within the strictly quantitative tradition of psychometrics, the most commonly collected sources of validity evidence are evidence for internal structure and criterion validity evidence. However, the inclusion of the three other types of validity evidence in the *Standards* represents a disciplinary position that these types are evidence are insufficient for validity. Since evidence of internal structure and relations to other variables can be assessed using strictly quantitative methods, while the other three forms of validity evidence almost always require qualitative data, this disciplinary shift towards additional forms of validity evidence amounts to a shift towards mixed methodology in validation.

In a discovery-based workflow, validation is an activity that seeks to learn new information about the overall measurement process. Some of this new information may arise from attempts to falsify claims about the measure. In justification-based workflows this new information either refutes the claim or leaves it intact. I suggest that a discovery-based workflow might engage in falsification-style reasoning but also seek to move beyond the binary implied in the process of falsification. An analogy might be drawn to the arbitrary alpha thresholds of significance testing: should a p-value of .06 convince us that the null hypothesis is a safe bet? Or should this information be used as an impetus to explore potential relationships more thoroughly? Moving towards the latter sort of reasoning puts us on the path to discovery. In the following discussion, I address the five areas of validity defined in the Standards with an emphasis on new information discovered about the measurement process during the investigation of each area.

5.2 Content Validity

Evidence of the validity of the content of a measure is collected in order to establish that the content of the instrument appropriately targets the construct. The content of the instrument should both fully represent the construct and omit any construct-irrelevant factors (Messick, 1995). In this study, validity evidence for the content of the measure is assembled from several sources and is combined into three knowledge representations. First, the evidentiary item map serves as one source of evidence, directly linking item content to real observations of the target population. This map can be consulted to trace the origins of each item. Second, the levels of the construct map are checked by a digital card sorting activity. Third, an iterative re-analysis of field notes is also compared to model output as triangulatory evidence about the prevalence of observed talk and behavior. Fourth, a tree generated by qualitative content analysis of alternative item task results is presented.

The Evidentiary Item Map is a knowledge representation which shows the link from ethnographic observations to items content. Excerpted summaries of field notes are included in one column, alongside records of other data sources which also mention the focal behavior (e.g. curricular materials), and the final text of the item included in the instrument. Any queries about item content can be answered by tracing back to the evidence employed for the writing of an item. If the appropriateness or adequacy of any item is doubted, an alternate, improved item can be written by consulting the excerpted ethnographic source data in the evidentiary item map. Evidentiary Item Maps are inspired by curriculum maps used for the professional creation of academic assessments (Jacobs, 2004), which are often employed as evidence of the validity of the content of instruments. While curriculum maps are

commonplace in educational assessment, to my knowledge, this practice has not been previously employed for the creation of behavioral and psychological measures in education.

Once model-derived estimates of ability were available, these estimates were compared to qualitative findings as another source of validity evidence. Since the items included in the instrument were all derived from ethnographic observation using the Evidentiary Item Map, instances of talk and behavior referenced in the items could be enumerated and categorized from field notes. Student-initiated talk and behavior were of particular interest, since the AHCS was ultimately formalized as a self-report measure. From the classroom field notes, tokens of student-initiated talk and behavior (n = 88) were indexed and coded according to which items they exemplified. For example, when a student was observed either answering a question about school work or talking about doing so, this was treated as an exemplar of the item "Answered a question asked by another student about school work," of which there were 5 student-initiated tokens among field notes. The estimated difficulty of a behavioral item is fundamentally a function of its frequency of endorsement in the population. As a result, we might expect items with lower difficulty to be observed more frequently in naturally occurring data and vice versa, an expectation which is treated as a testable hypothesis.

Student-generated behaviors matching the content of items were counted for all 22 items. In order to count as student-generated behavior, the focal topic must have been volunteered by the student, rather than any of the instructors.[26] These counts were compared to model-derived estimates of item difficulty using a simple linear regression. A statistically significant

[26] For example, for the item discussed in Chapter 4, "In the past two weeks, I have: Sought information about a professor of a class I want to take", only Context B was counted as student-initiated, since the other three tokens of this behavior pertain to instructor-initiated speech, either that of the faculty member or mentor.

negative relationship was found between the number of tokens of student-initiated behavior and its estimated difficulty ($p < .01$, $r^2 = .51$). Simply put, this means that the more frequently a student-initiated behavior occurred, the more likely participants were to endorse it on the survey. Due to the small sample size involved (n = 22), this regression is conducted purely for illustrative purposes.

Difficulty	# of Occurrences in Student-initiated Classroom Data
2.948	1
2.093	0
1.444	1
1.336	0
1.208	2
0.916	1
0.901	2
0.854	4
0.73	5
0.58	1
0.551	2
0.207	2
-0.065	1
-0.307	2
-0.362	2
-0.432	2
-0.471	7
-0.49	3
-0.984	3
-1.264	7
-1.625	7
-2.045	10

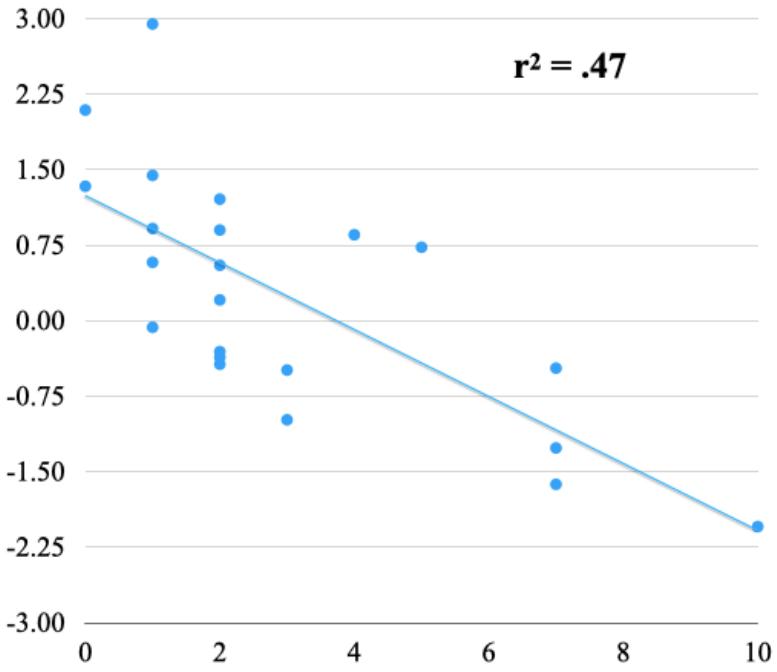


Figure 5.1: Item Difficulty and Number of Student-Initiated Occurrences in Classroom Data. The more frequently a behavior was recorded as observed or discussed in classroom ethnographic field notes, the less severe the item difficulty.

Finally, as part of a cognitive interview session, students were asked to contribute "additional items" for the AHCS. Three lists with different subsets of randomly selected

items were prepared for this task. The items that participants wrote down, as well as any additional items they spoke about during the exercise, were transcribed from the audio recording by a research assistant. The exercise yielded 81 suggestions for items. Many of these items semantically overlapped with the remaining items not shown to participants in this portion of the exercise. For example, seven of the 81 suggestions contributed by students were to contact instructors outside of class, either in office hours (6) or by email (1). Seven students suggested completing ungraded practice problems, and four students suggested attending the on-campus tutoring service - both of which were also items on the masked half of the AHCS. A qualitative content analysis (Schreier, 2012) was performed using the 81 item suggestions as cases. Qualitative content analysis (QCA) is a method for creating a dataset to categorize many distinct observations about text or images. The method is thorough and comes with a fully developed theoretical framework of its own. The basic procedure of QCA involves the following phases: select the data (e.g. interviews, internet comments, photos, videos, constructed responses), build a coding frame, code all the data, evaluate the coding procedure, and summarize findings. For this study, the strategy of progressive summarizing (Mayring, 2010) was employed to create categories for the items. In progressive summarizing, each case is summarized in more increasing general language until this general language also describes similar cases. For example, the following suggestions were all summarized as "group study activities":

> “Getting study rooms in the library with people in your class.”
>
> “Forming study groups.”
>
> “Made and joined a study group.”
>
> “Formed a class-specific study group.”

"Planning study sessions with friends at the library."

"Made a study group with friends."

"Worked on problems with friends who are in the same class."

As the categories from progressive summarization emerge, they are ordered hierarchically as subsets of larger categories. In this case, "independent study activities" also emerged as a category for another set of suggested items, leading to formation of a more general label "study activities" under which both "independent activities" and "group activities" were placed as mutually exclusive categories. Coding frames generated in QCA seek to structure and simplify content by adhering to the general principles that categories should be unidimensional, mutually exclusive, and exhaustive (i.e. each observation belongs to at least one category). The results of this QCA have been presented as a typology in Figure 5.2. All suggested items were able to be exhaustively subsumed into two major categories: *resources* and *self-management*.[27] *Resources* were *course resources*, such as textbooks and assignments, *campus resources*, such as on-campus tutoring, or student-selected *web resources*. Suggested items concerning self-management were subsumed into the three major categories of *well-being*, such as sleep and diet recommendations, *managing space*, such as managing noise and selecting an appropriate study location, and *managing time* - a category that required significant further subsetting. Suggested items pertaining to managing time included *scheduling* and *study activities*. *Scheduling* included recommendations about planning sequences of work and recuperation, while *study activities* included both *independent activities* and *group activities* meant to be pursued during study blocks. *Independent activities* were primarily *practice and review strategies*, such as taking practice

[27] The word "management" is employed here in the sense of handling and directing (etymologically, directing a horse, from Old French "manège"), rather than in the bureaucratic sense.

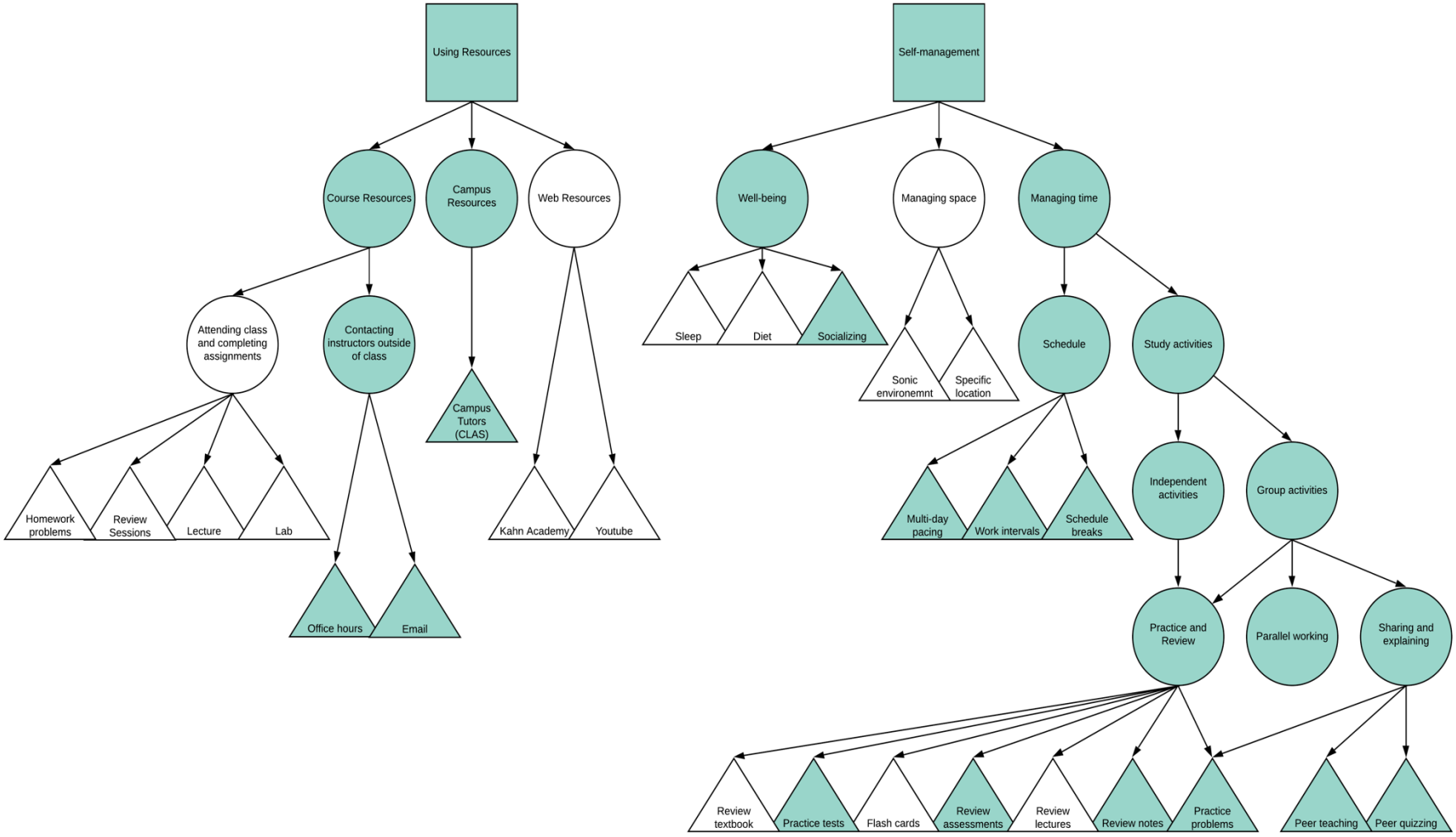


Figure 5.2: Construct Representation Tree derived from Qualitative Content Analysis. The green categories (circles) and specific habits (triangles) are those that are inluded in the AHCS, while white areas refer to areas of the construct not represented in this version of the scale. By this standard, this version of the AHCS had construct representation, although the white areas indicate potential areas for improvement.

tests and reviewing notes. *Group activities* included ways of *sharing and explaining*, such as quizzing peers, as well as *parallel working*, which involves working individually while co-present. Study groups may choose to engage in some of the same activities collectively as they would individually, although only *practice problems* was mentioned by participants in this study.

This visualization of the content of potential items is meant to showcase the intuitions of the target population about the manifestations of the construct. I have tentatively named this data representation a "Construct Representation Tree."[28] On its own, this tree useful for instrument development. By adding an additional layer of information, however, the tree can be used to assess the construct representativeness (Messick, 1995) of existing instruments. In this data representation, categories for which there are already semantically-similar items on the AHCS are colored green. Since the QCA coding frame is hierarchical in structure, if any lower node is colored then all its parent nodes are necessarily colored as well. For example, our scale includes an item about the use of on-campus tutoring and this leaf node is colored in, which means that the parent nodes of *campus resources* and *resources* are selected as well, since our scale contains at least one item pertaining to an instance of these categories.[29] This additional layer of information offers a visualization of the degree of construct representativeness at a glance, a powerful aid for measurement.

When combined with information derived from an item response theory (IRT) model, this procedure is suggestive of the kinds of new items that might be trialed to improve the

[28] One day, it is my hope that some bright mixed methods psychometrician will introduce a superior method. Anticipating this development, I suggest the devastating title "Constructs Don't Grow on Trees."

[29] If desired, additional color schemes might be developed to indicate a larger number of items or tasks pertaining to a single node, e.g. increasingly dark shades of green for more items on the instrument. For even greater precision, labels indicating the exact number of items could easily be placed in one corner of the triangular leaves, while the exact number of suggested items from the exercise could be placed in another corner.

existing instrument. As noted in the previous chapter, the Wright Map from the first administration of the AHCS showed that the instrument was well-targeted to most of the population except for a few participations with a low level of the trait, implying that a few low-difficulty items would improve the targeting of the instrument. Examining the construct representation tree, we see that most of the parent categories are at least somewhat represented, but the tree remains bare at some of the major nodes. The category of behaviors summarized as *attending class and completing assignments* is completely bare of items on our scale, and students were not queried about behaviors such as attending lecture and completing assignments. I suspect that this omission is likely due to limitations of my positionality: while developing items from the qualitative data at hand, I did not think to synthesize students' many remarks about engaging in such activities into a category for the construct map. To an academic gourmand like me, attending lecture and at least attempting homework are "minimum" academic habits that are too obvious to mention - I tacitly assumed that all students invested in passing notoriously challenging STEM courses would meet the minimum course requirements. Yet, combined with information from the model, it is clear that these behaviors would have made excellent candidates for low-difficulty items in the scale. They would also have provided useful information to the program head and contributed to our program theory. Future versions of the instrument should include items about these behaviors - which can be readily developed using field notes already assembled from the earlier phases - as well as items about managing space and web resources.

These three methods - creation of the Evidentiary Item Map, re-analysis of field notes in light of model-derived estimates, and creation of the Construct Representation Tree - offer strong validity evidence for the content of the measure. Such detailed accounts of the origins

of item content allow for the content to be evaluated in light of alternatives and tradeoffs in measurement. Most importantly, however, they permitted the discovery of additional aspects of the measurement process. Future iterations of the AHCS or derivative measures can use these knowledge representations as guidance when revisiting and perhaps making different choices.

5.3 Criterion validity

Criterion validity evidence for the instrument was established in three ways: 1) the ability of the measure to differentiate participants who took part in the intervention program from those who did not, 2) the ability of the measure to predict grades, and 3) comparison with a widely-used scale of a related construct. Assuming that participation in the BIOME program makes a difference in participants' academic behavior, as suggested by other studies of the program (Wilton, 2019), then our measure of academic habit complexity should be able to differentiate program participants. Latent regression can be used to compare estimates of ability between participants and non-participants while accounting for measurement error and ensuring that intervention effects can be interpreted in the same unit as item difficulties and person abilities (Adams et al., 1997; De Boeck & Wilson, 2004; Lu, Thomas, & Zumbo, 2005). A latent regression was performed to estimate the effect of the intervention. This is formalized as a two-level model:

Level 1: $Log[\Pr(X = 1)] = \theta - \delta$

Level 2: $\theta = \beta_1 * Program\_Treat + \epsilon_p$

Where the probability of endorsing an item is a function of θ (person ability), and δ (item difficulty). β is the estimated effect of the program and *Program_Treat* is an indicator variable that takes a binary value. The top equation is simply a restatement of the Rasch model, while the bottom equation is a regression using the latent variable of person ability (θ) to predict whether a student was enrolled in the intervention program, plus some allowance for error (ϵ) at the person level.

Results of the latent regression showed a statistically significant difference between participants in the program and those in the general population of biology majors, accounting for measurement error. An average participant in the program had a latent ability .30 logits higher (se = .14 logits) on academic habit complexity than non-participants. This is comparable to an independent samples t-test which treats the habit complexity as an observed rather than latent variable, which likewise proves to be significant, $t(268) = -2.35$, $p = .02$. In sum, BIOME participants scored 5% higher on the AHCS than non-participants, and this difference was statistically significant.

The second form of criterion validity evidence to establish was the relationship between academic habit complexity and performance in a difficult STEM class taken by first-year biology majors, Chemistry 1a, final grades for which were obtained for all students in the study via administrative request. Final grades in chemistry were not available at the time of construct selection, construct definition, instrumentation, or data collection, eliminating a potential source of experimenter bias. Additional indicators of participant characteristics, such as prior academic preparation and demographics, are also available to increase the precision of the estimates and check for interactions.

After the model-fitting stage, the academic habit complexity scale was used to predict final grades in a difficult STEM course taken by first-year biology students, Chemistry 1a. Model-derived estimates of person ability (Θ) were computed for each participant, ranging from -2.91 to 2.47 logits with a mean of .05 logits.[30] A multiple linear regression (n = 263) reveals that the estimate of academic habit complexity scale (AHCS) is a statistically significant predictor of chemistry grades, both in a simple linear regression and after controlling for URM status, gender, SAT Verbal, and Pell-grant status (which were non-significant predictors) and SAT math and high school GPA. The final model, including only the significant covariates, consisted of SAT math and high school GPA, which function here as rough indicators of prior academic preparation, and the AHCS, which significantly predicted chemistry grades at the $p < .005$ level, $b = .177$, $p = .001$. In practical terms, this means that for every four academic habits students reported, we can expect Chemistry grades to be half a letter grade higher on average. Simply put, students with very high Chemistry grades attested to many more academic habits than students in the lowest range. To assess the degree to which Habit Complexity improved the quality of evidence provided by the model to predict chemistry grades, a Bayesian linear regression was performed using SAT Math and HS GPA as part of the null model and then hierarchically adding Habit Complexity. The Bayesian linear regression of Habit Complexity on Chemistry grades yielded a BF_{10} of 27.63, meaning that the data were 27 times more likely under the alternative model than under the null model. A Bayes Factor of this magnitude is traditionally interpreted as strong evidence for the alternative hypothesis (Kass & Raftery, 1995; Wagenmakers et al., 2016). In

[30] Alternatively, a raw score from 0-22 points yields similar results. It is a feature of the Rasch model that raw scores are "sufficient statistics."

other words, the inclusion of Habit Complexity considerably sharpened the model's overall ability to predict chemistry grades to an $r^2 = .44$.

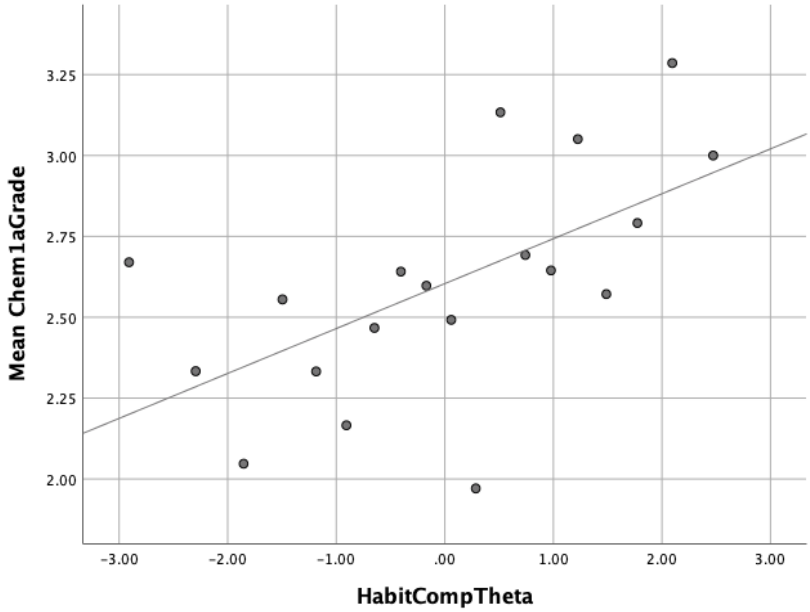


Figure 5.3: Simple Scatter of Means of Academic Habit Complexity and Chemistry Grades.

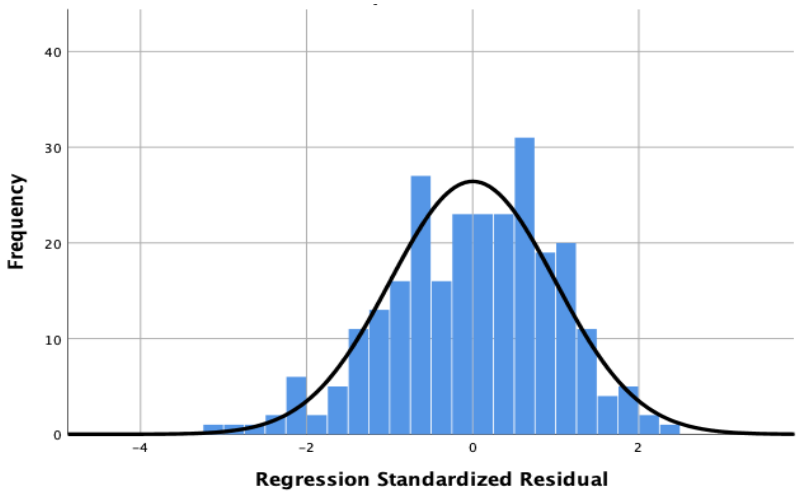


Figure 5.4: Standardized Residuals of Regression in Figure 5.3.

A second, previously published scale of academic habits ("academic integration", AI), originally developed for used by the National Center for Education Statistics (Wine et al., 2011) and adapted for numerous studies of higher education (e.g. Flynn, 2014, Xu, 2018, Solanki et al. 2019), was included in the study by the PI, and was used as one benchmark against which to assess the quality of the AHCS. Little information is available about the genesis of this AI scale, although the use of Cronbach's alpha in previous studies as the sole indicator of measure quality suggests that this scale was created within a Classical Test Theory framework. Both the relationship of the AHCS to chemistry grades and the benchmarking against a previously published scale were double blinded from the perspective of the development of the AHCS, since neither chemistry grades not the content or structure of the second scale were known at the time of implementation. I hypothesized that the both the ACHS and the AI would predict grades in chemistry, that the AI would be related to hours spent studying STEM subjects, and that both scales would be able to detect differences participants who took part in the intervention and those who did not.

Several trials suggest that AHCS was a more appropriate measure for the target population than AI. First, AI did not predict chemistry grades in our sample ($p = .49$), as did the AHCS (above). In no combination of main effects and interaction effects with the selected covariates from the previous model was AI statistically significant. Second, the design of the AI scale indicates that it was meant to measure the *frequency* with which the five selected academic behaviors were accomplished - what is called "habit strength" in the psychological literature (Verplanklen & Aarts, 1999). Curiously, however, the total score from the AI scale did not correlate with students' self-reported hours spent studying STEM subjects. Habit complexity as measured by the AHCS, however, was significantly related to

self-reported hours spent studying STEM subjects. Third, the AHCS was not able to detect differences in the behavior of participants in the focal program and those who were not, as did the AHCS (see latent regression, above). In sum, the AHCS performed as well as expected, but the AI scale did not. This is unsurprising given that the AI scale focused on only on four academic habits – all concerning contact with faculty and staff[31] – and lacks evidence of validity for first-year students (See Chapter 2). As Flynn (2014) notes, correlations among these items are weak for first-year students (ranging from .23-.38), suggesting that these items may not measure a common construct for first-year students.

These three forms of evidence attest to the criterion validity of the instrument by showing that it predicts another program outcome and that it outperforms a commonly-used predecessor. The successful prediction of chemistry grades suggests that the instrument describes aspects of academic behavior that are helpful for, but not determinate of, real-world success.

5.4 Response Process

Validity evidence based on the response is assembled to demonstrate a causal link between the construct and the instrument (Borsboom et al., 2004). Ideally, the individual's experience with the construct causes them to form a memory, which is then queried by the instrument, leading to an appropriate response on the instrument. There are many potential events that can obstruct this process, leading to a failure of measurement. For example, a participant may have actually engaged in a targeted behavior or possess the requisite skill

[31] The 4-item instrument used by the National Center for Education Statistics (NCES) asks about the frequency of 1) formal contacts with faculty outside of class, 2) informal contacts with faculty members outside of class, 3) meetings with academic advisors, and 4) use of study groups (Flynn, 2014).

queried on the instrument, but then be unable to recall this knowledge or ability at the time of administration. Alternatively, a participant may be able to recall the relevant information but choose to provide an invalid response to the instrument by engaging in deception. Events of this kind disrupt the causal chain of events leading from the construct to the instrument, and are thus considered to be threats to the validity of the response process.

Cognitive interviews are widely considered the method of choice for collecting evidence about the response process (Willis, 2004; 2015). In classic cognitive interviews, participants are briefly trained to offer a "think-aloud" narrative of their thoughts, then asked to complete the instrument in a safe and confidential setting. This stream-of-consciousness account is composed of both concurrent and retrospective narrations of thoughts (Taylor & Dionne, 2000). While cognitive interviews do not purport to penetrate the deepest recesses of the mind, they offer insight into a layer of thinking that is usually inhibited in normal conversation. Indeed, the method is an outgrowth of forensic investigatory techniques developed to facilitate the recall of information of which witnesses may not have been directly aware, such as the physical description of a bystander in a bank robbery (Fisher & Geiselman, 1992). Since this classic cognitive interview method was popularized for psychometrics, additional modules have been used to augment it, including category sorting and other tasks.

In this study, one round of cognitive interviews was conducted to clarify any issues in comprehension of the instrument content (n = 7) and a second round of cognitive interviews (n = 20) was focused on other aspects of the response process. In the first round of think-aloud cognitive interviews, participants demonstrated that all items were understood as intended. Small syntactic edits were performed at this stage, but the semantic content of the

items remained the same. Participants produced the prototypical and expected response processes to all items, through the comprehension, retrieval, judgement, and response stages (Tourangeau et al., 2000).

In the second round of cognitive interviews (n = 20) participants were asked to narrate the events of a day when they recently spent a considerable amount of time on school work. This was treated as an open-ended conversation and participants were encouraged to give as much detail as possible. Students' narratives about a recent day spent engaging in school were compared to the content of items on the AHCS. This portion of the data attests to the causal link between the construct of academic habit complexity and students' ability to recall and verbally encode information about the construct. Personalized one-day timelines were created from each of the student responses for purposes of comparison with the instrument. The behaviors enumerated by students overlapped significantly with the behaviors targeted by items on the AHCS, such as completing practice problems, studying with peers, and scheduling breaks. Behaviors mentioned by students that lacked a direct analogue on the instrument were often more general statements about academic behaviors that were more clearly specified on the instrument, such as "I studied in the library" - whereas on the instrument "studying" is broken down into many separate behaviors. This was not judged as a threat to the validity of the response process, since students were able to recall many specific behaviors as well.

5.5 Internal Structure

Validity evidence for the internal structure of the measure typically includes an analysis of item fit and reliability. The bulk of this evidence is collected during the model fitting stage

(Chapter 4: Instrumentation), and will only be briefly summarized here. First, the data fit the hypothesized model on the first attempt, without adjusting any parameters of the model post hoc. No items were deleted, no additional factor structures (e.g. bifactor models) were imposed, and no error terms were allowed to correlate to improve model fit. Infit mean-squared statistics are visualized in the Fit Web (Figure 4.1). Second, the fit of the data to the model suggest that the construct can be adequately modeled as unidimensional. A PCA of standardized residuals did not recover any unexplained substructures. In addition, the unidimensional model was found to have superior fit (BIC) to a multidimensional Rasch model. Third, empirical item difficulties conformed to the expectations embodies in the construct. Namely, academic behaviors that students undertake on their own are easier to complete than activities involving peers, and activities involving individuals of higher social status (instructors, upperclassmen) are the most difficult of all. To take a sampling of these three categories: working practice problems had a difficulty of -2.05 logits, group study had a difficulty of -0.47 logits, and contacting an instructor outside of class had a difficulty of 1.44 logits.

Fourth, more than 90% of participants fit the hypothesized model, suggesting that the model provides an excellent general description of the construct for this population. Persons were approximately equally likely to fit the model regardless of demographic group or program membership. Person-separation reliability was .77, indicating that the scale can be used to differentiate between multiple, qualitatively distinct groups of participants (Linacre, 2018). Fifth, person-item targeting was generally successful - the mean item difficulty was not far about the mean person-ability in this population. However, several easier items would have helped the measure to cover the lower range of person ability. Sixth, items were

checked for differential item functioning by observed groups of gender and URM status. Some items showed significant DIF given these observed groups.

Ethnographic observations, focus groups, cognitive interviews, did not reveal any additional information about how gender or URM status might affect these items - like Antoine de St. Exupéry, we find ourselves flying blindly in a storm with only our instruments to guide us. Since the implications for "fairness" in this measurement are minimal (e.g. the AHCS is not used to make high stakes decisions that might lead to systemic inequalities), I suggest that the issue of DIF can be treated here as an area of interest for future iterations of the measure. If DIF can be reproduced later for the same items and in the same directions, then deleting items may be warranted. Whereas deleting items would be the safe option in a justification-based workflow, since it eliminates a potential line of critical attack, a discovery-based workflow would seem to favor collecting further information about an unusual finding before deciding that it amounts to a problem with the instrument.

5.6 Consequential Validity

Since instruments can only be valid for a particular interpretation and use (Messick, 1989), the *Standards* require the collection of evidence about the interpretation and use of instruments in context. To the discredit of the respective fields, this requirement is rarely met in actual practice for academic assessment (Oliveri et al., 2018) and even less for psychological assessment. In keeping with my overall purpose in creating workflow for the development of an educational measure, I address this problem by providing guidance about how to formalize the representation of evidence about the consequences of measurement.

The consequences of measurement, I suggest, can be schematized along three dimensions: the scope of consequences, the likelihood of consequences, and the positive or negative valence of consequences. The scope of consequences begins with particular individuals and causally ripples out through increasingly distant strata of social organization. The ripple metaphor is only accurate to a point however, since some consequences of measurement become greater at greater distance from the individual, producing an inverted ripple that grows in size instead of dissipating. The likelihood of consequences causally extending to each stratum depends on the specific characteristics of the measurement activity. Some projects are so local in scope that macrosystem consequences are nearly inconceivable, while other projects, such as PISA, are designed to advance international goals. Practitioners typically know which of these situations they are in. Consequences can also be characterized as positive, negative, or mixed at each level. Some instruments cause individuals psychological discomfort, while others offer a chance for reflection on behavior and attitudes. Unexpectedly, the growing popularity of online self-assessments of all types has revealed a large popular demand for instruments that purport to guide self-exploration and self-labeling. The use of some measurement practices, such as value-added measurement in education, has also caused harm from the mesosystem upwards (Rothstein, 2012; Darling-Hammond et al., 2012, AERA, 2015). Other measurement practices clearly represent known tradeoffs of positive and negative consequences. This is indeed the normative premise behind the computation of classification accuracy - false positives and false negatives should be minimized to avoid the consequences of denying or wasting valuable services. The requirement that consequences be assessed requires practitioner to make ethical judgments about the valence of these consequences.

When these three dimensions - scope, likelihood, and valence - are considered together, overlaps emerge that set priorities for the research workflow. It may be difficult and costly for most individual researchers or practitioners to gather evidence for every level of the scope of consequences. Effort should focus on gathering evidence at levels of social organization where consequences are most likely. Measurement used to make only local decisions requires less evidence about possible distant consequences than do measurement activities at the regional or national level. The overlap between level of social organization and likelihood of consequences dictate where effort and expense should be dedicated at this stage.

To capture these three dimensions in a single visualization, I propose a Metrological Bronfenbrenner Chart. Known to many social scientists, the Bronfenbrenner ecological model shows the potential interactions of an individual with increasingly distant strata of social organization (Bronfenbrenner, 1979). The individual at the center of a series of concentric circles interacts with a microsystem (e.g. family), a mesosystem (e.g. school or workplace), ecosystem, (e.g. extended family, neighborhood), and macrosystem (e.g. nation, ideology). The consequences of measurement can be located on the scope of this continuum. For example, taking a self-scored quiz at home affects only the individual, while taking a classroom test can affect every layer from the individual to the exosystem. Other instances of measurement, such as the census, have much greater impacts on the macrosystem than on any lower level. The Bronfennbrenner chart serves as the basis of the knowledge representation that will allow for the addition of other dimensions. I propose that the likelihood of consequences be layered on the Bronfenbrenner chart as an ordinal score from 0-3:

0 - impossible

1 - unlikely, < 20% chance

2 - likely, 21-79% chance,

3 - highly likely, > 80% chance

A national census, for example, might score a 1 on the individual level but a 3 on the macrosystem level. These likelihoods are not meant to be assigned empirically, but rather represent an important prediction by researchers and practitioners about the impact of their work. When in doubt, it is advisable to study the real-world impact of similar measures to determine the likelihood of impacts at various levels. For example, the designers of the ACT could have made a reasonable prediction about the consequences of their test by analogy to the SAT, which preceded the ACT by several decades. The likelihood scores are located in the 9 o'clock position, 90 degrees from the level labels, which are in the 12 o'clock position. To save space, the chart can thus be truncated to a quarter-slice of the full circle.

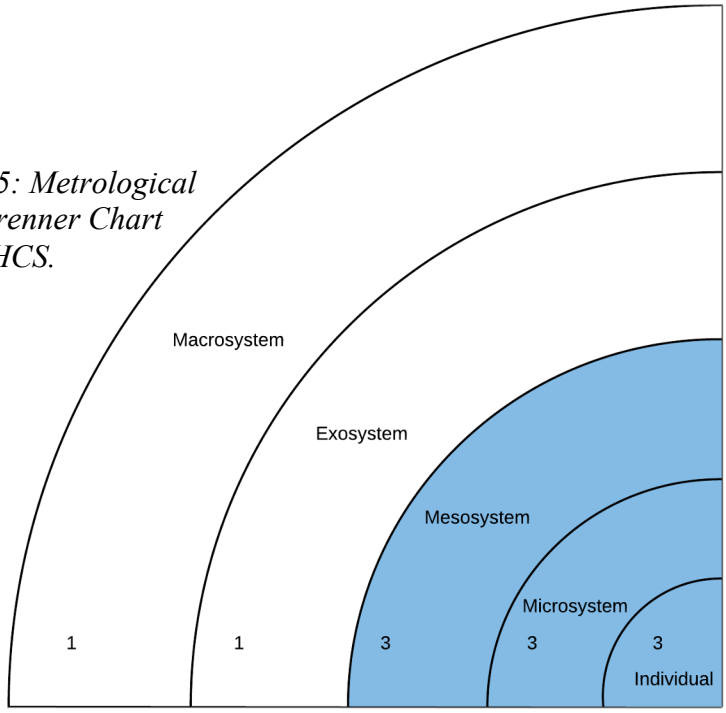
The third dimension to be added to the data representation is the valence of consequences. Are consequences of measurement known to be positive, negative, or mixed at each level of social organization? Let the color blue stand for positive consequences, the color red stand for negative consequences, and the color purple stand for a mixture of positive and negative consequences. A particularly noxious and unfair standardized test, for example, might have known negative consequences at multiple levels of social organization relative to a particular individual: red for the individual (personal discomfort), red for her microsystem (family disappointment), red for her school (facilitates harsh austerity), but

purple for the macrosystem (it enables the identification of problems but also causes serious social dysfunction).

The fourth dimension to be added to our knowledge representation is our information state about consequences at each level. When information is available about consequences at this level, the layer can be colored in. When no information is available at a particular level, this layer should be entirely white. The metrological Bronfenbrenner charts for measures that are used repeatedly can be updated in subsequent publications in which the measures are used. Alternatively, the chart can be updated via the creator or vendor's website, as subsequent evidence about the consequences of use becomes available.

Using these rules, the metrological Bronfenbrenner chart for the AHCS is given likelihood scores of 3 at the individual level, 1 at the microsystem, 3 at the mesosystem and exosystem levels, and 1 at the macrosystem level. Correspondingly, evidence for the consequences of the AHCS should be presented for those levels of social organization.

Figure 5.5: Metrological Bronfenbrenner Chart for the AHCS.



Consequences at the individual level of the AHCS are judged to be positive based on the available evidence. The investment of students taking part in the AHCS is a favorable one for two reasons: first, the demands placed on their time, attention, and emotional state are deliberately minimized, and second, participants may receive a monetary reward. Little is sacrificed and there is a modest potential gain. A subset of students participating in the cognitive interviews were asked identify whether any items on the scale might cause "hesitation" for any reason. None of the reasons identified by participants indicated that any items contained sensitive content. Additionally, participants indicated that the exercise of completing the AHCS was "interesting" and that it "made me think about what I'm doing in school right now." Before viewing the scale, cognitive interview participants were asked to narrate the actions they took on a day recent day when they completed "a lot of school work." These narrative responses contained information that was remarkably similar to the content of the AHCS, such as accounts of completing practice problems, asking friends questions during study sessions, scheduling social activities at the end of the day, and going to campus-based mental health services. No outward manifestations of discomfort, such as hesitations, refusals, or other interactional trouble, are apparent in the interviews. Participants were quick to answer our query about their school-related activities on a previous day, often simply listing activities in quick succession. The willingness of students to spontaneously produce detailed narratives of construct-relevant behaviors serves as evidence that these behaviors are not taboo and that the items are not sensitive for participants in this population.

Consequences of measurement at the microsystem level might consist of effects on family or roommates. Since the AHCS is a form of confidential self-report, and does not require reporting on the behavior of these other people, no consequences are anticipated at

this level of social organization. (Compare this to a survey of attitudes and behaviors of family members or romantic partners, for example.)

The mesosystem and exosystem levels are much more affected by the use of AHCS, since these are the levels at which schooling is situated. In particular, the measure was employed as part of a process-improvement evaluation (Chen, 1996) of a program. Results from the measure have informed the development of the program in three ways: 1) by indicating areas where curriculum might be improved to intervene on the difficulty of some items for observed groups, 2) indicating that there is a significant difference between students who participated in the program and those who did not in the achievement of these goals, 3) offering support for hypotheses embodied in the logic model of the program (Newcomer et al., 2015). Further explanation about the role that the AHCS has played in the development of the program are detailed elsewhere (Clairmont, Katz, Wilton, in progress). As this was part of a process-improvement evaluation, all evaluation data are marshaled for the positive goal of program improvement (Chen, 1996). Thus, the consequences of measurement for the program are expected to be positive for program managers, program staff, and program beneficiaries (Greene, 2000).

The Metrological Bronfennbrenner Chart matches well with recent developments in validity theory in education. For example, Haertel (2013) presents an expanded view of validation that focuses on seven purposes of evaluation ranging from "instructional guidance" to "shaping perceptions." Fulfilling these purposes requires consequences to occur at different levels of social organization, from the individual to the macrosystem. These consequences can be located on the Bronfenbrenner chart and reported using the data representation described here. Evidence supporting historical challenges to constructs such as

IQ (Rindermann et al., 2020), which are argued to cause harm at the macrosystem level rather than at the individual level (e.g. in clinical settings), might also be more clearly represented using this chart.

The Metrological Bronfenbrenner chart is meant to summarize evidence about consequences without necessarily endorsing a single view of the relationship between consequences and validity. Messick's view of consequences, for example, has been critiqued as both too liberal (Borsboom, Mellenbergh, & van Heerden, 2004; Popham, 1999) and too restrictive (Cronbach, 1988, Inoue, 2009) in the consequences that can be considered. More ink has certainly been spilled in the debate about what sorts of consequences should be considered than what these consequences actually are for real assessments. Given that consequences are a heavily contested area in measurement, I suggest that at least reporting them in a formal structure is a best practice, even knowing that readers and practitioners are free to apply normative judgments about the consequences depicted in the chart.

5.7 Validity Argument

The validity evidence collected for a measure can be formalized as an inductive argument (Kane, 1992). The argument-based approach to validity holds that particular inferences in this argument can be evaluated based on the strength of the evidence for each step of the argument, as well as the plausibility of warrants for the inferences. Putting all of the pieces of the validity argument together allows for the appraisal of the entire inferential process.

The validity argument for the measure of academic habit complexity begins with a theory of academic habit complexity. The evidence that gives rise to this theory comes from the

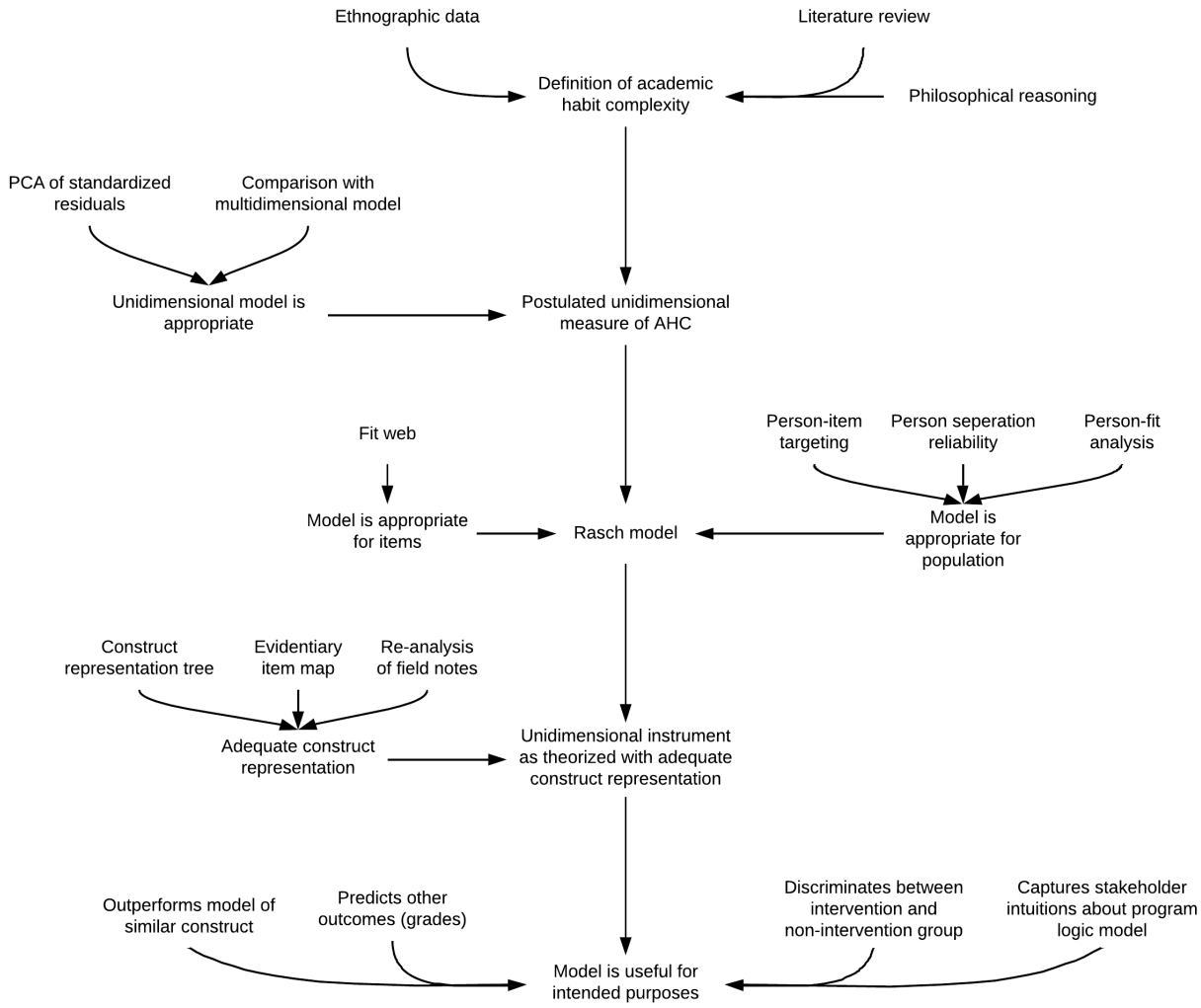


Figure 5.6: Validity Argument for the AHCS.

ethnographic study that began this project (Chapter 2). Previous research and philosophical reasoning is also brought in at this phase (Chapter 3. This theory leads to the postulation of a unidimensional measure of academic habit complexity (Chapter 4). Additional evidence for instrumentation comes from the evidentiary item map (Chapter 4), which is also based on the original ethnographic research (Chapter 2). Evidence for the content of the instrument is also provided by the Construct Representation Tree (Chapter 5). This postulation of a

unidimensional measure allows sets up the next step in the argument, which is that Rasch analysis is applicable. Four pieces of evidence are marshaled for this key step in the argument. First, a PCA of standardized residuals and a comparison with a multidimensional model are conducted to check that a unidimensional model is appropriate (Chapter 4). Second, person-item separation is checked to determine whether the model is able to distinguish participants with different levels of the construct (Chapter 4). Third, person-item targeting is analyzed to determine the extent to which the measure has an appropriate level and range of difficulty for the population (Chapter 4). Fourth, person fit analysis is analyzed to determine what proportion of participants fit the hypothesized model. In sum, this evidence provides backing for the key claim that, for this population, the Academic Habit Complexity Scale is unidimensional as theorized, with comprehensive construct representation and meaningful levels. The scaffold of inferences involved in the validation of this instrument are depicted in Figure 6.5.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*.

American Educational Research Association. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. Educational Researcher, 44(8), 448-452.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel Item Response Models: An Approach to Errors in Variables Regression. *Journal of Educational and Behavioral Statistics*, 22(1), 47–76.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, 111(4), 1061–1071.

Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design*. Cambridge, MA: Harvard University Press.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer and H. Braun

(Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.
Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, 93(6), 8-15.
De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer Science & Business Media.
Fisher, R. P., & Geiselman, R. E. (1992). *Memory-enhancing techniques for investigative interviewing: The cognitive interview*. Springfield, IL: Thomas.
Flynn, D. (2014). Baccalaureate attainment of college students at 4-year institutions as a function of student engagement behaviors: Social and academic student engagement behaviors matter. *Research in Higher Education*, 55(5), 467-493.
Greene, J. C. (2000). Understanding social programs through evaluation. *Handbook of qualitative research*, 2, 981-1000.
Haertel, E. (2013). How is testing supposed to improve schooling?. *Measurement: interdisciplinary research and perspectives*, 11(1-2), 1-18.
Hagquist, C., & Andrich, D. (2015). Determinants of artificial DIF-a study based on simulated polytomous data. Psychological Test and Assessment Modeling (Vol. 57).
Inoue, A. B. (2009). The technology of writing assessment and racial validity. In C. Schreiner (Ed.), *Handbook of research on assessment technologies, methods, and applications in higher education*. Hershey, PA: Information Science Reference.
Jacobs, H. H. (2004). *Getting Results with Curriculum Mapping*. Association for Supervision and Curriculum Development, Alexandria, VA.
Jaeger, R. M. (1987). Two decades of revolution in educational measurement!?. *Educational Measurement: Issues and Practice*.
Kane, M.T. (1992) An argument-based approach to validity. *Psychological Bulletin*, 112, 527- 535.
Kass, R. E., & Raftery, A. E. (1995). Bayes factors. Journal of the American Statistical Association, 90, 773–795.
Linacre, J. M. (1998). Detecting multidimensionality: which residual data-type works best?. *Journal of outcome measurement*, 2, 266-283.
Linacre, J. M. (2018) A user's guide to Winsteps Ministep Rasch-model computer programs 2018. Chicago: Winsteps.com.
Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological reports*, 3(3), 635-694.
Lu, I. R. R., Thomas, D. R., & Zumbo, B. D. (2005). Embedding IRT in Structural Equation Models: A Comparison With Regression Based on IRT Scores. *Structural Equation Modeling: A Multidisciplinary Journal*, 12(2), 263–277.
Mayring, Philipp (2010). *Qualitative Inhaltsanalyse. Grundlagen und Techniken [Qualitative content analysis. Basics and techniques]* (11th rev. ed.). Weinheim: Beltz.
Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, 18(2), 5-11.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9), 741.

Newton, P. E., & Baird, J. A. (2016). The great validity debate.

Oliveri, M. E., Rutkowski, D., & Rutkowski, L. (2018). Bridging Validity and Evaluation to Match International Large-Scale Assessment Claims and Country Aims. ETS Research Report Series, 2018(1), 1-9.

Popham, W. J. (1999). Where large scale educational assessment is heading and why it shouldn't. *Educational Measurement: Issues and Practice*, 18,13-17.

Rosas, S. R., & Ridings, J. W. (2017). The use of concept mapping in measurement development and evaluation: Application and future directions. *Evaluation and Program Planning*, 60, 265–276.

Rindermann, H., Becker, D., & Coyle, T. R. (2020). Survey of expert opinion on intelligence: Intelligence research, experts' background, controversial issues, and the media. *Intelligence*, 78, 101406.

Rothstein, J. (2012). Effects of value-added policies. *Focus+*, 29(2).

Schreier, M. (2012). *Qualitative content analysis in practice*. Sage publications.

Solanki, S., McPartlan, P., Xu, D., & Sato, B. K. (2019). Success with EASE: Who benefits from a STEM learning community?. *PloS one*, 14(3).

Taylor, K. L., & Dionne, J. P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology*, 92(3), 413.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Verplanken, B., & Aarts, H. (1999). Habit, attitude, and planned behaviour: is habit an empty construct or an interesting case of goal-directed automaticity?. *European review of social psychology*, 10(1), 101-134.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, 25(3).

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design.* Sage Publications.

Willis, G. B. (2015). *Analysis of the cognitive interview in questionnaire design*. Oxford University Press.

Wilton, M., Gonzalez-Niño, E., McPartlan, P., Terner, Z., Christoffersen, R. E., & Rothman, J. H. (2019). Improving Academic Performance, Belonging, and Retention through Increasing Structure of an Introductory Biology Course. *CBE—Life Sciences Education*, 18(4), ar53.

Wine, J., Janson, N., & Wheeless, S. (2011). 2004/09 Beginning Postsecondary Students Longitudinal Study (BPS: 04/09). Full-Scale Methodology Report. NCES 2012-246. *National Center for Education Statistics*.

Xu, D., Solanki, S., McPartlan, P., & Sato, B. (2018). EASEing students into college: The impact of multidimensional support for underprepared students. *Educational Researcher*, 47(7), 435-450.

Zumbo, B. (2007). Three generations of DIF analysis: Considering where it has been, where it is now and where it is going. *Language Assessment Quarterly*, 42(2), 223–333.

Chapter 6

Conclusion

In this conclusion, I revisit the four normative strands of the workflow enumerated in the introduction and make explicit arguments for each in light of the case study in the preceding chapters.

First, it was argued that measure development should embrace a discovery-based approach instead of a justificatory approach. What is at issue here is a difficult question for scientific activity in general: what conditions warrant exploration or confirmation? Obviously both exploration and confirmation are important and there is little point in attempting to sideline one of these elements. They are distinct modes of inquiry, however, and it is no small matter to decide when each is needed. To grasp the difficulty of this decision, imagine that researchers are in possession of a magical epistemic map of their domain: it is darkened in the areas where more exploration would be required to uncover the basic facts about that region and it is bright in the areas in which much is already known. In the areas in which much is known, confirmatory efforts are appropriate, but a discovery-based approach is more appropriate in the darkened areas. Research questions in those darkened areas should contain fewer prespecifications and presumptions, and the explorer should be more prepared to encounter surprises and unforeseen impasses that might warp the meaning of any methods tentatively employed. I argue that when developing a measure in a region of epistemic darkness, discovery-based methods are much more appropriate. "Caesar had had the entire region reconnoitered" (*Civil War*, 9.68) is fairly typical remark when reading of his campaigns - commentators have noted that Caesar's choice of *regiones* versus *loca* ("places") is important, since it indicated that "Caesar had explored far more than straight ahead and

was now going to take advantage of his more comprehensive understanding of the terrain" (Raaflaub & Strassler, 2017). Measuring the same "known" constructs again and again, without exploration of the adjacent unknowns, risks dividing the world of educational phenomena into isolated *loca*. Discovery-based approaches confront the issue of context and unknowns directly, mapping more of the region in which we find familiar places. Ending the thought experiment and returning to a world in which there is no enchanted epistemic map, it is all too apparent that the few known *loca* of clarity do not offer enough reassurance to embrace justificatory reasoning as a default practice in measure development. This realization was arrived at by Lee Cronbach (1975) after the failure of a decade-long research program seeking to discover nomothetic generalizations in education, leading him to make the following recommendation:

> Instead of making generalization the ruling consideration in our research, I suggest that we reverse our priorities. An observer collecting data in one particular situation is in a position to appraise; a practice or proposition in that setting, observing effects in context. In trying to describe and account for what happened, he will give attention to whatever variables were controlled, but he will give equally careful attention to uncontrolled conditions, to personal characteristics, and to events that occurred during treatment and measurement. As he goes from situation to situation, his first task is to describe and interpret the effect anew in each locale, perhaps taking into account factors unique to that locale of series of events (cf. Geertz, 1973, chap. 1, on "thick description"). As results accumulate, a person who seeks understanding will do his best to trace how the uncontrolled factors could have caused local departures from the modal effect. That is, generalization comes late, and the exception is taken as seriously as the rule.... When we give proper weight to local conditions, any generalization is a working hypothesis, not a conclusion (pp.124-125).

The substantial differences between contexts brought Cronbach to the realization that local conditions needed to be accounted for prior to attempting measurement. His reference to Geertz suggests exactly how thorough such an accounting would need to be. I would like to

think that something like the workflow proposed in the preceding chapters of this dissertation would have suited the old master.

The second normative strand motivating the present workflow is the injunction to adopt radical transparency. This injunction can be explained in terms of the psychometric concept of the validity argument. Validity arguments lay out the inferences employed in the development and use of measures. These arguments are the formal representation of a process of reasoning about measurement from evidence. This is why, as Kane (1992) says, inferences require warrants and backing: these terms are names for the formalizations of the role of evidence in the argument. As Kane (1992) has noted, arguments that are not articulated cannot be evaluated. The idea that we should see *radical* transparency is thus only a slight extension of Kane's original argument. Here *radical transparency* denotes a total disclosure of evidence and findings, a "warts and all" portrayal of the measurement process without airbrushing, whitewashing, or other metaphors for pretense. Where problems are encountered, they should be disclosed. Where items or persons do not fit the expected model, this should be investigated. Where participants dispute the premise of a question, their objections should be explored. Among researchers, the primary point of disagreement with this normative strand seems to be that the aesthetic standards of some subfields of educational and psychological inquiry do not permit such openness. But what seems more likely, that transparency is wrong or that these aesthetic standards are misguided and unproductive?

The third normative strand in the workflow is the deliberate integration of philosophical exploration of the construct into the workflow. This means that constructs should no longer be treated naively as though they are objects to be found lying about. Continuing this ignores

decades of patient anthropological and sociological work on the practice of doing science, which emphasizes scientific constructs are hybrids: partly natural and partly socially constructed, partly discovered and partly made (Latour, 2012). Emphasizing discovery to the exclusion of construction allows for naive realist and operationalist understandings of constructs to proliferate. As Alexandrova (2018) has explained, such "evidential subjectivism" leaves us unable to explain the relevance and completeness of construct definitions (p.135). After a time, reified constructs circulate in studies like so many unattended pets wandering through a crowded party. A cry goes up, "Big 5 Conscientiousness has bitten me!" but its owners are absent at the scene of the crime. Such is the social life of a construct undisciplined by philosophical definition and critique. In practice, philosophical reasoning about constructs does not raise separate issues about measurement, but rather hastens the discussion of issues that are likely to arise in time as generalizations are attempted and implicit meanings unfold through repeated use.

The fourth normative strand motivating the workflow was the importance of integrating qualitative methods at each stage of the measure development process. The argument for doing this runs as follows. For the sake of clarity, let us take as given the four-step measure development workflow introduced in this dissertation. At least two logical stages of measure development, the construct selection and construct definition process, do not appear to be amenable to statistical methods. Truly inductive quantitative techniques are not available for construct selection, since these techniques rely on too high of a level of methodological pre-specification (i.e. items and constructs have already been selected; Wilson, 2004). Construct definition cannot proceed entirely using statistical methods without falling prey to the fallacy of operationalism. The next two logical stages - instrumentation and validation - do not

require qualitative methods in the same way that the first two steps do. Rather than being a necessity, qualitative methods are optimal for these final two phases. For example, in the instrumentation phase, employing focus groups and cognitive interviews in the development of items is obviously better than not doing so, since it offers a chance to conduct needed revisions. Similarly, in the validation stage, it is obviously better to collect qualitative evidence of the validity of the measure's content than not to do so, and it is difficult to imagine how statistical evidence alone might serve this purpose. Thus, integrating qualitative methods into the workflow of measure development is either necessary or optimal at every step. To a certain extent, the foregoing argument relies on the distinctiveness of qualitative and quantitative methods - a view with I do not endorse fully - but the difficulty of drawing a bright line through the messy territory of "qualculation" (Cochoy, 2002) would itself seem to lend support to the spirit of the main argument. This fourth normative strand amplifies the power of the other three by stating the context in which the previous three can be exercised. The use of qualitative methods at each phase of measurement provides a stage on which discovery, transparency, and philosophical reasoning can fully play their parts.

Rather than simply keeping these normative strands "in mind", why should we instantiate them in a workflow? As Tal (2016) argues, the evidentiary strength of measurement claims lies in their epistemic security. This security is derived from using precautions and strategies which reduce the number of possible scenarios in which the measurement claim might be incorrect (Staley, 2012). The contribution of the present project to add to the stock of precautions and strategies which bolster epistemic security in measurement. In other words, I suggest that, while measurement outcomes are not mere observations, assembling observational data at right junctures can enhance the epistemic security of measurement. One

purpose of the “instruments of the mind”, Bacon says, is to supply “cautions” - a workflow can be such an instrument.

References

Cochoy, F. (2002). *Une sociologie du packaging ou l'âne de Buridan face au marché: Les emballages et le choix du consommateur*. Presses Universitaires de France.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. American psychologist, 30(2), 116.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., ... & Weiner, S. S. (1980). *Toward reform of program evaluation* (p. 3). San Francisco: Jossey-Bass.

Kane, M.T. (1992) An argument-based approach to validity. *Psychological Bulletin*, 112, 527- 535.

Kuhn, T. S. (1996). *The structure of scientific revolutions (3rd ed.)*. University of Chicago press.

Latour, B. (2012). *We have never been modern*. Harvard university press.

Raaflaub, K. A., & Strassler, R. B. (Eds.) (2017). *The Landmark Julius Caesar: The Complete Works: Gallic War, Civil War, Alexandrian War, African War, and Spanish War: in One Volume, with Maps, Annotations, Appendices, and Encyclopedic Index*. Pantheon.

Appendix I: Group Interview Protocol

Introduction

The purpose of this interview is understanding BIOME and improving it. Your participation in this interview is confidential, and you may withdraw at any time. Everyone will be given pseudonyms, and your instructor will not listen to this recording. Do you consent to being recorded?

Interview Prompts

How has school been going so far?

In general, what are your biggest challenges since coming to UCSB?

How did you hear about BIOME the first time? Walk me through your decision to join BIOME.

What do you think about the other people in BIOME?

We're three weeks into the quarter now. How much do you value BIOME, from not at all to somewhat to a lot? Why?

Please write down your personal pros and cons of participating in BIOME. After this, add anything in the middle ground too.

Has BIOME changed your views of anything in particular, perhaps about the university, about the major, or about your own outlook?

Do you want to revise anything you've said so far?

Appendix II – Evidentiary Item Map

Item	*Field Note(s)*	*Documents*	*Illustrative Context*
In the past two weeks, I have:			
Attended a tutoring session, such as CLAS	1, 2, 4, 6, 9, 10	MH (mentor handbook)	Mika is experimenting with different study strategies. "Have you been going to CLAS" Steve asks. "It doesn't help." Mika responds, softly. The students and Steve all jump in to suggest that she switch sections of CLAS and go to a different tutor. They compare notes on the different tutors. "Try the drop-in hours" Steve advises. "Have you gone to office hours?" Steve asks. "It's scary!" Mika protests. Sonya jumps in excitedly, "Go! Go! Go! It's not bad!"
Studied with another student in my class	1, 3, 4	MH	The instructor asks for advice from people who have already figured out some academic strategies that work for them. A muscular guy in the front of class advises students to work together with others in their classes. The instructor spent ten minutes making sure that students were sharing contact information. "I'll see you in class tomorrow" a young woman says to a new acquaintance afterward. Speaking of his study group in college, the instructor summarizes "There was solidarity in doing the work together."
Asked another student a question about my school work	3, 9		"This time I studied with another person and it was way better" Mika says, reflecting on her last test, "I felt comfortable asking questions and they would explain it."
Talked to a TA or faculty member outside of class,	2, 3, 5, 6, 10	MH	The instructor argues that attending office hours as crucial to success. A quick show of hands reveals that half the students have already been to office hours at least once - not bad for week three. A guest speaker, who is a professor, talks about "the importance of getting to know your

including office hours			professors" and says that "you need to foster relationships with your professors" recounting the situation of being asked to write a letter of recommendation for a student she cannot remember. She invites students to attend her office hours even if they are not in her class.
Planned my social time around my study schedule	1, 2, 3, 5, 10	MH	Homework for this week includes an exercise in which students will estimate how many hours they spend on various activities and then compare this to an actual time diary. "Figure out how much free time you have in that bottom box," The instructor says, explaining how to subtract planned work time from students' 168-hour week. A student states that his goal is to push himself as hard as he can to get into med school, but the tutor reminds him to have fun too. "Reward yourselves for getting things done," the instructor tells the students, after suggesting that they avoid returning home in the evening until they have accomplished everyone on a daily checklist. Raúl begins his check-in with an earnest update, "I think college is easier that I thought. I thought it was gonna be impossible." "Just manage your time, study, and have fun" Steve reassures him. Steve concurs, and advises T3 to study well in advance and then take the night off before the test.
Emailed a TA or faculty member directly	1, 9, 10	CW, MH	When a student mentions a class he is in, the TA responds "My roommate TAs that class," and offers her help if the student needs anything, "just send me an email if you need anything from her." The instructor offers "If you want to get into [a specific lab on campus], send me an email and I'll help you craft a message to [the PI]."
Sought information about a professor of a class I want to take	1, 3, 5, 10	MH	The instructor advises students "never to walk into a class on the first day without knowing how that class operates" explaining that other people are a valuable source of information. Raúl explains that he has realized that students who get lower grades tend to review professors

			poorly. Sonya agrees and says as a result, she no longer trusts ratemyprofessor.com as a source of information. “Take what people say about a class with a grain of salt,” Steve affirms. “I’m scared about biostats,” a student says. Steve gives lots of advice for how to plan one’s coursework. During the growth mindset group discussion period, he warns freshmen to watch out for a pileup of difficult course in the second year, telling them that things even out after that. He even suggests taking physics during study abroad to skirt the local physics department. Today Steve ended up giving advice about math, chemistry, and physics classes. One of the mentors describes his major emphasis, CSS, and The instructor asks him to give lots of details. Again, students are silently attentive. Later, when asked about his personal study strategies, Steve says to the class “Find out about the professor and what’s going to be important to them.”
Sought information about an internship or lab position	3, 5, 6, 9	MH	Brigit gives a student advice about how to get letters of recommendation, get into research labs, and meet graduate students. At the tail end of the whole-class discussion, students recommend that the hypothetical student take an internship or other topical classes. Steve explains how to look online for labs to join, “What are they studying? How are they studying it? You’re not gonna understand everything but look at the titles.”
Taught a class concept to another student	7, 9		“If you can’t teach someone how to do something, you don’t know how to do it!” The instructor says, urging students to study with peers. Raúl talks to the table about studying with someone who struggles more than he does with chemistry. He says that he doesn’t mind teaching them.

Sought information about an academic society or academically-oriented Greek organization	3		After class, The instructor coaches a freshman about how to interview well for a medical fraternity.
Made use of campus-based support programs such as CAPS, Student Health, or the AS Food Bank	4	MH	Mika was sick recently, but realized that she knew very little about medical care at UCSB. “Where am I supposed to get medicine?” she recalled thinking.
Avoided spending time with people who keep me from getting my work done	3, 4		Izzy and Sonya get into a conversation about students (all men, in this discussion) who they find distracting and difficult to be in class with. Steve listens attentively and then advises them to try to ignore them, since there isn’t much to do about this. Referring to her study plan, Sonya explains “I made a distinction between studying by myself and studying with friends, cause with friends you’re only studying part of the time.”
Worked on practice questions that won't be graded	2-10	MH	Each table computes an average of the number of chemistry practice problems done: 11.5, 4.3, 9.8, etc.
Created and followed a study schedule	2, 3, 6, 7, 9, 10	MH	The mentor at Table 3 repeats the “college is your job now” mantra, spelling out the 9-5 schedule in detail. The instructor’s standard for work is the 40-hour work week. This session, Steve asks T3 whether they’ve done their 40 hours.

			The instructor explains the danger of "putting out sequential fires" by neglecting all other subjects while cramming. Raul discusses how he wound up failing his math midterm. "It was the same day as the chem midterm, so I didn't study for math" he tells the table. Steve reacts visibly, straightening in his chair and slightly raising his voice. He is not shouting, but he is speaking with some authority. "See, this is a time-management problem," Steve pronounces.
Practiced for tests or quizzes using a timer	2, 6	MH	Guest Speaker: "What does it mean when you run out of time?" Student: "You can't do the problems fast enough." Guest speaker: "And how do you learn how to do problems faster?" Students: "Practice?" "Time yourself?" "Practice tests." The black woman with cornrows at T2 tells the class that she uses a timer when doing practice tests before exams.
Delayed a reward for myself until after I met my academic goals for the day	3, 7, 9		"Reward yourselves for getting things done," The instructor tells the students, just after suggesting that they avoid returning home in the evening until they have accomplished everyone on a daily checklist. "I promise your efficiency will go up if you do that," he says. Brigit recommends setting up personal rewards for meeting one's academic goals. Sonya listens to classical music during her breaks to refresh herself. Steve returns to the theme of rewards, "You gotta work for those breaks - it's your job now."
Started studying for a test or quiz more than three days in advance	3, 6, 10		The instructor asks students who have just taken a quiz what advice they would have given themselves 24 hours before. The student responses emphasize studying more planning, "reviewing material at least a week in advance." "What did you learn from that midterm?" The instructor asks. He asks who studied the night before the midterm, then 2, 3, 4, & 5 nights out

			from the midterm. “The idea is to prepare for the exam as far ahead as possible.” “OK, who has actually started studying for finals?” Steve asks T3, a week ahead of time.
Reworked problems that I missed on previous assignments	6, 9, 10		Only about 1/3 of the class completed both parts of the exam wrapper, and The instructor reminds the rest to do this at the end of class. Steve emphasizes the value of returning to previously missed problems. “If you didn’t do an exam wrapper because ‘I don’t want to think about it, it makes me sad’ then you’re setting yourself up to miss those items on a cumuluative final.”
Marked problems or concepts to study again later	4, 10		In whole-class discussion, Sonya advises marking the book problems that you’re getting wrong and coming back to them. “Make use of old exams and practice exams, go back to notes, to the book, the ones you got wrong, find out where to focus” the T1 mentor offers.
Taken a step back from my work to judge my overall understanding	8, 9		Jenna reveals that she more than doubled her score from the previous quiz, inciting praise from T3. “So, what did you change?” Steve asks (calling back to Nov 1st). “I studied by myself more… I would write down a solution and then come back to it later,” Jenna answers. “I figured out my weakness, which is lab, so I started going to a different TAs office hours before I had to turn those [labs] in” Sonya says to T3. She says she is learning to monitor her performance in different parts of the class. However, she admits, she didn’t make time to adequately prepare well for the last chemistry quiz. “If you had a midterm, how did it go and how did you study for it” Steve begins the conversation about preparing for finals. The TA suggests that students carefully schedule avoid spending too much time without a break so

			that they can "recognize when your studying is no longer efficient."
Planned ahead to take a relaxing break before a test or quiz	5, 7	MH	During whole-class discussion, Izzy advises taking a nap before major tests. Steve concurs, and advises T3 to study well in advance and then take the night off before the test. During the class discussion about stress in the final weeks of the quarter, one of the mentors (T5) recommends simply taking a day off to recuperate.
Double-checked my work before turning it in	7		When Jenna completed her exam wrapper, she realized that she forgot to bubble about 20% of her exam answers onto the proper sheet, many of which were correct. She says that if she had only checked her work, this wouldn't have happened. Steve consoles her for this mistake, emphasizing that this means she actually knew more than she thought she did.
Note: MH = Mentor Handbook, CW = course website			

Appendix III: Item text, Item Difficulty, and Differential Item Functioning

Item Difficulty, Fit, and DIF in Academic Habit Complexity Scale

	Delta	Infit MNSQ	DIF Favoring
Attended a tutoring session, such as CLAS	-0.31	1.111	URM (1.1)
Studied with another student in my class	-0.47	0.996	
Talked to a TA or faculty member outside of class, such as office hours	1.44	0.937	URM (.50)
Planned my social time around my study schedule	-1.26	0.967	nURM (.66)
Emailed a TA or faculty member directly	1.34	1.006	
Asked another student a question about school work	-1.63	0.872	
Answered a question asked by another student about school work	0.73	1.010	
Sought information about a professor of a class I want to take	2.09	1.057	M (.64), nURM (.61)
Sought information about an internship or lab position	-0.98	0.876	
Participated in activities with an academic society or academically-oriented sorority or fraternity	2.95	0.987	F (.48), nURM (.48)
Made use of campus-based support programs such as Student Health, CAPS, or the AS Food Bank	1.21	1.110	F (.59), URM (.85)
Chosen not to spend time with people who keep me from getting my work done	0.55	1.028	

Worked on practice questions that won't be graded	-2.05	0.928	F (.51)
Created and followed a study schedule	0.85	1.110	F (.77)
Practiced for tests or quizzes using a timer	0.58	1.101	M (.62)
Delayed a reward for myself until after I met my academic goals for the day	0.21	0.981	
Started studying for a test or quiz more than three days in advance	-0.49	0.952	
Reworked problems that I missed on previous assignments	-0.07	0.918	
Marked problems or concepts to study again later	-0.43	0.932	F (.89)
Taken a step back from my work to judge my overall understanding	0.90	0.947	M (.80)
Planned ahead to take a relaxing break before a test or quiz	0.92	1.089	M (.61)
Double-checked my work before turning it in	-0.36	1.048	M (.55), nURM (.7)

Notes. Sample size adjusted critical range for MNSQ statistics is 84-1.16. M = men, F = women, URM = underrepresented minority, nURM = Whites and Asians, reference category. The magnitude of significant DIF approaching .5 logits and greater is reported in logits beside the group that the DIF favors.