〔コミュニケーション情報学篇〕
【論文】

Teaching formal measurement in the humanities and the specific case for a foundational understanding of self-report inventories

Ian ISEMONGER

要旨（Abstract）

In this paper, the fundamental contours of conceptual understanding which students in the humanities should have with respect to the critique and use of self-report inventories are outlined. The general importance of a critical approach to measurement is outlined as the broad context for this paper, and then the discussion is restricted to one of the most commonly-encountered measurement tools, the self-report inventory. The paper discusses key conceptual notions for pedagogical and curricular focus. It is argued that students should receive a conceptual understanding of the principles behind these instruments which assists with a more critical posture, and avoids a naive presumption concerning the integrity of the data they generate.

キーワード（Keywords）：measurement, pedagogy, humanities, education

Introduction

A signature feature of modern society is that we are all measured in some form or other every day of our lives, and that this measurement affects us in ways which are often opaque but nonetheless impactful. In terms of our increasing interaction with the Internet as consumers, whether for the purposes of information search and retrieval, commercial activity, social networking or any number of other types of interaction, our actions in this domain are constantly measured by interested parties, and this measurement impacts the scope and nature of subsequent interaction, as systems adapt to inferences made about us on the basis of this ongoing measurement. In education, and to the largest degree in higher education, students evaluate the performance of their teachers through questionnaires or surveys, though the validity of these evaluations might sometimes be questionable (Shevlin et al., 2000; Spooren et al., 2013), while, in turn, teachers evaluate students' performance on the material taught through a variety of tests. Teachers are often required to self-evaluate their own performance on institutionally-implemented systems as part of professional practice, and there are a growing number of instruments for teachers to self-evaluate outside of institutional implementation, and as part of personal professional development. Outside of education, students entering the workforce will continue to experience, and sometimes execute, depending on their role, all manner of evaluations and tests including job performance evaluations, customer satisfaction evaluations, and market evaluations. Even in the context of applying for a position itself, students may encounter a battery of tests and evaluations in the form of aptitude tests, personality tests, leadership tests and so forth (Wolf, 2002). These examples represent only a fraction of the multitude of ways human behavior is constantly measured in modern and complex industrial societies, and the manner in

受付日：　2017年11月９日
受理日：　2017年11月16日

which decision making and human action is increasingly rooted in this measurement.

While it would be true to say that we have always lived under some sort of regime of assessment, what distinguishes the above examples are the formality and explicitness of the assessment. While in the past, for example, a business manager might have made hiring and promotion decisions using the more casual framework of intuitive and commonsense assessment of daily performance, more modern assessment systems require explicit and quantifiable assessment which is capable of being used in the open justification of decisions. In the past, decisions made by a manager may have been evidence-based, but were so predominantly in the mind of the manager and his/her casual observations. Today, however, new forms of explicit and data-driven assessment are underpinned by method, and are there to provide accountability for decisions, and often public accountability, with respect to any number of other agendas such as non-discriminatory practices and so forth. This explicitness of modern assessment is amplified by its overwhelming presence in modern life, and one could continue to list examples in all manner of contexts.

From a humanities point of view there are two critical issues which present in view of this state of affairs; a state of affairs I would argue to hold true mostly in industrially-advanced countries. The first, which is worth more than a mere noting, but which is outside the scope of this paper, is how this measurement impacts on the scope and parameters of human action and our experience of living. It is the humane question of what it means to “be” in such a world. How, for example, does the potentially overwhelming experience of constantly being measured in some form or other affect happiness and the general human condition, and how does it constrain or alter our behavior? These sorts of questions are truly questions for which the humanities is uniquely positioned to engage. The second is how to give students the analytical tools to engage critically with measurement wherever they may encounter it. The imperative for this is constantly increasing as the range of objects subjected to rigorous and quantifiable measurement continues to expand. The rise of digital humanities, for example, which includes as one component of its agenda, a quantifiable approach to the domain of literary text (Hockey, 2004; Fitzpatrick, 2012), indicates another extension of the hand of objective measurement into an area, i.e. literary text, traditionally dominated by the scholarly approach. Measurement is necessarily the issue of quantifying observations, and while there are clearly field specializations within the humanities and especially the associated social sciences (like psychology, sociology, and economics) which rely fundamentally on quantifiable data, and therefore formal measurement, the emphasis given to understanding the theory behind such measurement is often insufficient in the humanities more broadly. And yet, given the pervasiveness of measurement to the modern human experience, and its extension even into disciplines traditionally not associated with it at all, like the humanities, it is something that all should understand, not just some. Indeed, there is a case to be made for such awareness being a life skill of sorts.

Addressing this deficit comprehensively is a task of quite substantial breadth. For example, students need to understand how measurement of online behavior and interaction occurs in the digital world and how they function as product, rather than consumer, in this process (Solon, 2011; Goodson, 2012). They need to understand, as students, how the regime of achievement or proficiency testing they perform under works, and how it affects them; not to mention what the strengths and limitations of such testing are. They need to, as students in the humanities, understand how new methods related to so-called “big data” provide new eyes for old problems (Moretti, 2007), with the quantitative approach to text in digital humanities, referred to above, being a subset of this. They need also to understand how they may be assessed as they enter the world of work. Because measurement has become so pervasive, there is indeed much to engage with. Outlining a pedagogical and curricular direction for this, therefore, is a considerable endeavor which exceeds the scope of a single paper.

One form of measurement, however, which is particularly important to understand, and which is increasingly present in modern life, concerns psychometric instruments which claim to measure stable aspects of mind. These instruments fall under the general category, or field specialization, of mental measurement, but would exclude achievement and proficiency tests which also constitute a form of mental measurement. Sometimes they are commercially published, and used under copyright, usually for a fee. Most times, however, they are unpublished of themselves, but appear in published research literature, in journals for example, as measurement tools used for academic research. It is these unpublished instruments which students are likely to encounter, or adopt, as part of their own research, and the number of such instruments in the literature is, quite simply, enormous. Of this type of instrumentation in the literature, one subcategory is even more likely to be encountered or adopted by students, and this is the self-report inventory. These instruments are of particular importance given the frequency with which they tend to appear in student theses, whether as part of the literature review and the generation of data for the findings of previous research reported therein, or as instruments they choose to adopt from the literature to generate their own data. Outside of academia itself, they are also important given the frequency with which they may be encountered as part of various personnel selection processes, and in particular, as part of candidate selection for job applications and employee placement.

Self-Report Inventories

The American Psychological Association has published, and continues to publish on an ongoing basis, a series of volumes which index the unpublished psychometric instruments appearing in published literature. This series is entitled the *APA Directory of Unpublished Experimental Mental Measures* and currently Volume 9 (Goldman & Mitchell, 2008) is the latest contribution. There are an extraordinary number of instruments indexed in the volume and this is from only 36 of the top journals in the fields covered by the volume. Twenty four categories are represented which include achievement, adjustment, development, motivation, values, vocational interest, and trait measurement, to name just a few. Outside of these 36 journals which are the source for the volume, there would be many more instruments to be found. Of these, a large proportion would fall under the category of the self-report inventory, and as stated above, this category of instrumentation is the most likely to be encountered, or used, by students. Their distinguishing feature is the fact that, as the name suggests, administration involves the respondent self-reporting, usually in some form of expression of the extent to which a particular statement applies to them. The natural question which follows, therefore, concerns why these type of instruments are so pervasive.

The Pervasiveness of the Self-Report Inventory

The most important reason for the extensive use of the self-report inventory in research is that it short-cuts the significant challenge, and often tedium and invasiveness, of field observation, on the one hand, and the technical, and often expensive, challenge of controlled laboratory observation on the other hand. Field observation also has the limitation, which is a practical one, of being non-conducive to the collection of large datasets. All mental measurement requires measurement of behaviors which are hypothesized to operationalize the mental constructs that the researcher is interested in. After all, we cannot measure the mind directly. We can only measure behaviors which operationally reflect aspects of mind. The obvious way to do this is to go into the field and observe the operational behaviors of interest, with these observations being rigorous, rather than casual, in some specified way. While this represents the most direct way to measure for the purposes of data collection, and the most ecologically valid way, the fact remains that it is difficult and time consuming, and we can only watch a few participants at a time. In the laboratory case, which is essential if the proposed research requires experimental design for causal inferences, it is easier to make observations of more people in less time, but there are still significant material challenges in setting up the technical aspects of laboratory control; not to mention the cost of these aspects, and the cost of paying participants for their valuable time to come and be placed under observation.

The self-report inventory assists with overcoming these obstacles, although it has to be said that this applies mostly to field observation and any other kind of research (e.g. descriptive or correlational research) where the research design does not necessarily involve the laboratory. This is because laboratory observation is usually put in service to the experimental method, i.e. when there is a need for strong causal inferences, and the self-report inventory is not generally associated with controlled observation under these types of design. It may be used in some ancillary way, for example to assist with the assignment of participants to experimental and control groups, if not by simple random assignment, but it cannot usually substitute for the measurement required by the experimental method itself where behaviors are experimentally manipulated and then objectively measured. The self-report inventory, therefore, overcomes many of the difficulties of collecting observations in the field, and in getting large datasets, and does so without being overly intrusive into the daily lives of participants.

Outside of research practice, the self-report inventory is also widely adopted in various forms of professional practice, for example, counselling and career guidance. In these cases, this form of instrumentation is adopted because it offers, again, a fast track for the professional to gain information about the person they are involved with professionally assisting. Additionally, it facilitates the gathering of such information in a relatively systematic way, and if the instrument is well-established with an evidence-based background for its usefulness, there are also usually clear paths to the interpretation of the information provided.

The Design of a Self-Report Instrument for Mental Measurement

Before outlining the design of a self-report instrument for mental measurement, it is important in a preliminary way, and for the purposes of clarity, to distinguish self-report instruments for mental measurement from opinion surveys, teacher evaluations, course evaluations and other general survey tools. These latter types of instruments measure opinions, which are obviously held in the mind, and in this sense constitute some form of mental measurement, but which nonetheless are about fairly contingent issues out in the world. If we take opinion surveys about political leadership, for example, a very popular and often-encountered form of survey, the purpose of these is to measure what the populace thinks about the performance of a government, a particular leader, a candidate for election, a candidate political party and so forth, and importantly, at a particular point in time. The results we get from these surveys can fluctuate considerably depending on a variety of factors such as the economic environment, the changing policies of political parties, changes in leaders and candidates and so forth. Similarly, a survey of student opinion on teacher performance can also fluctuate, and is a contingent form of survey. The results can vary according to the teacher, the course, the changing methods of a teacher, the level of the students, expected grade, prior interest in the course, and so on (for evidence related to many of these see Wachtel, 1998), although validity and reliability of results from these kinds of evaluations has also been supported (Marsh, 1984; Marsh, 2007). In this kind of way, therefore, opinions held temporarily in the minds of citizens or students, or whomever, are used, essentially, to take a look at aspects of the world they perceive whether this be the political landscape, the classroom and teacher, or any other number of other areas. This is not the same as dedicated mental measurement associated with trait theory where the purpose of the survey is essentially to measure relatively stable aspects of mind which influence behavior, in one way or another, relatively predictably. Instruments which measure intelligence, aptitude, personality, personal adjustment, social adjustment, and other similar constructs we presume to be relatively stable, fall into this category. These instruments attempt to provide data which is descriptive of relatively stable aspects of mind.

The differences between these two classes of instrument are often expressed in the structure or design of the instrument. Typically, though not always, for example, in a teacher performance survey there will be one item per point of survey interest. So, if one point of survey interest were about students' opinions on the teacher's punctuality, there would be one statement in this regard, something like this: "The teacher was always punctual." Students would then reply on, for example, a Likert scale, where they would elect to endorse one of the following with respect to the statement: "Strongly Agree," "Agree," "Neither Agree nor Disagree," "Disagree," or "Strongly Disagree." This single item would then comprise the data point for student opinion about the teacher's punctuality, and the purpose of the item is often to gather data about the teacher and his/her behavior as much as it is to gauge what is going on in the minds of students. It is important to note, that in some cases some items might group around an aspect of behavior, for example, how much a teacher cares about students. There may be three or four survey items which target caring behavior in the teacher; however, these locate themselves around the behavior of the teacher, and not some property of mind of the student doing the evaluation. Although we do not presume the score awarded to be entirely a property of teacher performance because the rater (in this case the students) influences the score awarded, it is the performance of the teacher which primarily occupies the purpose of the whole endeavor. Now aside from the critical posture we should have on this kind of survey, like whether the popularity of the teacher biases the information students provide about his/her punctuality, the key point is that mental measurement would depart from this form of survey in two important ways. First, in mental measurement properties of mind linked to the respondent are the target of the survey, rather than properties of the outside world (i.e. the performance of another party such as a teacher) which just happen to be held in mind; i.e. in the mind of the rater, or student doing the evaluation. Second, in mental measurement there will typically be more than one data point targeting a particular property of mind.

Turning to the first point, that properties of mind are the target of a survey in this form of mental measurement, the underlying assumption here is that these properties are relatively stable and this is associated with trait theory—a paradigmatic approach to psychological research with a long pedigree. In this sense the properties of mind we are typically interested in are not transiently held opinions about states of affairs in the world (classroom, teacher, political landscape, leader or anything otherwise) and therefore contingent, but rather, and on the contrary are about fairly (note, not absolutely) non-contingent and stable aspects of mind which will play out and determine behavior in a manner which follows a predictable pattern. We are also not using mind, or opinions held in mind, as a proxy to measure something else, but rather are measuring mind itself. Turning to the second point, that there will typically be more than one data point per property of mind taking our interest, this is best illustrated with an example. If one were interested in measuring the propensity for an individual to be either extraverted or introverted, one would be measuring here two constructs of mind, namely, extraversion and introversion(1), and we would presume these to be relatively stable properties of mental disposition and predictive of general behavior. We cannot measure these constructs directly, in the sense that we cannot see extraversion or introversion in the mind of a person, all we can see is behaviors which indicate an extraverted or introverted mental disposition. We would also presume that these constructs of mind cause certain behavior, and not the other way around. Thus if we want to measure these constructs we would have to seek out behaviors which count for them, or in the language of formal method, behaviors which operationalize or indicate them.

The important point here is that in mental measurement, we would seek out more than one behavior indicating the construct which takes our interest, so as to fully reflect it. So, for instance, and with respect to extraversion, we would seek out behaviors like propensity to accept invitations to parties, propensity to speak in groups, propensity to avoid working alone and so on. Because we presume the construct to represent a trait, and therefore cause a general pattern of behavior rather than just one specific behavior, we need to measure a number of behaviors which fall under the circumscription of the construct. If we measure too few behaviors, we fail to fully express the behavioral bandwidth of the construct from an operational point of view. In the field, therefore, we might observe a number of behaviors like the ones identified immediately above which indicate extraversion, but this would be difficult. In the case of a self-report instrument, we would shortcut this difficulty by presenting respondents with a number of statements which directly reference these behaviors, asking them to report the extent to which these behaviors are true, or not true, about themselves, with the extent often being expressed as a frequency. There may also be, in the instrument, a number of other items where respondents are asked to report on the extent of their own behavior with respect to another construct, in this case for example, introversion, and this is because an instrument usually measures more than one construct of interest. Critically, the conceptual separation of the two constructs needs to be reflected in an empirical separation between the sets of scores designed to measure the respective constructs. In other words, we presume scores on one construct to be associated with each other, and dissociated from scores on another construct.

Thus aside from the problem of how truthfully respondents might respond to these statements, which is a general problem for all self-report instruments (mental measurement or otherwise), the foremost point is that the instrument will have a design which proceeds well beyond a simple list of statements. Each statement in the instrument will be associated with the construct, or latent, which we presume to be causing the response on it. So if we were to have an instrument which measured extraversion and introversion (two constructs), we might have ten self-report items on the instrument with five intended to measure extraversion and five intended to measure introversion. This association of items with constructs in an instrument, in a formal and designed way, opens up a range of methods attendant to the formal assessment of how well the instrument actually works, and how secure the data it generates can be assumed to be. As such, the association of items with constructs also represents the foundation of what is to be taught in any educational response to helping students critically engage with these types of instruments.

Pedagogical Aspects of a Conceptual Understanding of Self-Report Instruments

From a pedagogical point of view, and here I mean pedagogy with respect to students in higher education who need a critical understanding of these kinds of instruments, the essential starting point is a satisfactory grasp of the distinction between the latent and the observable; with the observable variously being referred to as the operationalization, indicator, or variable. The latent, in juxtaposition to the observable, is distinguished by being precisely not that; i.e. precisely not observable. It cannot be directly seen and measured. It corresponds with a construct or concept, and is by nature abstract. It is part of the conceptual tooling for theoretical reasoning. So we might theorize that extraversion (Construct A) tends to correlate with social adjustment (Construct B). Neither Construct A nor Construct B can be directly seen in the mind, we can only see behaviors that might indicate them. However, we do not want to couch our theoretical reasoning in the multiplicity of behaviors which might result from them, because this is messy and forfeits the abstraction which is characteristic of theorizing; and it is abstraction precisely which makes theorizing powerful through force of generalization. At the same time, we do want to test our theorizing of the relationship (in this case positive correlation between Construct A and Construct B) in observable reality; and that reality is precisely the multiplicity of behaviors which we do not want to clutter our theoretical reasoning. It is exactly the need to talk, think and theorize in constructs and abstractions, and the equally important need to have these constructs and abstractions rooted and tested in an observable reality, which give self-report instruments their structure.

From a fundamental understanding of the above distinctions between the latent and the observable, students will have the conceptual frame of reference to begin appreciating what the term “measurement model” means when comprehending how a self-report instrument for mental measurement works, and how the scoring regime for the instrument is associated with this measurement model. This is because the measurement model is essentially defined by how the observables are claimed to map onto the latents, and the word “claimed” here is important because this mapping is fundamentally something which is first claimed, and then subsequently tested for actually being the case or, indeed, not being the case. Thus, the measurement model and the associated scoring regime are the very roots of a critical approach when using such an instrument, and when considering how another researcher may have used it. It is in service to verifying the plausibility of the measurement model that a variety of technical/statistical procedures are carried out (see below for discussion).

A measurement model is essentially hierarchical with individual items on the instrument (which are the observables) hypothesized to be subordinate to one, and only one, latent/construct. Put differently, this means that an item, or observable, should only measure one latent/construct and not any other present in the overall instrument, and important statistical methods are used to test the extent to which this ideal is, in practice, the case. Returning to the extraversion and introversion constructs for further example, five items would be subordinated to the extraversion construct and five to the introversion construct. None of the five items measuring extraversion should also be measuring introversion and vice versa. In short, items need to be, in principle, exclusive indicators of a construct. This allows for the conceptual separation between two constructs, like extraversion and introversion, to be equally expressed at the operational or observable level, and this is a critical issue to be grasped by students. If the observables (items) do not provide separate measurement of relevant constructs (empirical separation), then the constructs we want to analytically or theoretically separate for the purposes of theory are, in fact, not testable because empirical separation is not provided.

It was briefly stated above that the scoring regime for an instrument is associated with the measurement model, and there needs to be evidence for this model. It is imperative for students to understand, and this is a suitable stepping off point for instruction, that the scoring regime for an instrument does not exist because the person making it decided it to be so, or it should not be anyway. On the contrary, the scoring regime should be rooted in an evidence-based model, and the issue of evidence here is critical. Students are often very familiar with the notion of gathering data as evidence for some sort of research claim, whether it is evidence for a claim about which they are reading or one they want to make themselves. They are also familiar with the idea that gathering data requires measuring something, and that the data from the measurements are the basis for evidence and empirical inference. What they are often not so familiar with is the notion that there should be evidence for the measuring itself. Giving students a grasp of how evidence is accumulated for measurement with respect to self-report instruments, first and foremost requires giving them the conceptual understandings outlined above with respect to the latent and the observable, and the manner in which the measurement model relates to the correspondence, or lack thereof, between these two. In turn, the scoring regime for a self-report instrument depends on the measurement model, and the extent to which this model has an empirical reality. This empirical reality and some of the methods associated with verifying it, and about which students should be familiar, are outlined below. However, from a pedagogical point of view, it could be concluded that the first steps in heightening students' awareness of evidence-based measurement are essentially foundational and start with a conceptual understanding of the structure of these instruments. From a practical point of view, students' engagement with the structure of instruments can be accelerated by always forcing the question with respect to any given instrument they may be dealing with: "What is the measurement model for the instrument?" If students are operating under any kind of naive assumption that items comprise a simple list of discrete, single-point measurements, to be interpreted item by item, then the above question will start the process of a more comprehensive awareness. In addition to this, if students' are aware at some level that certain items group together in some way, the above question has the advantage of encouraging them to make their awareness more explicit, thus enhancing their critical posture.

Essential Statistical Methods for a Critical Approach to Self-Report Instruments

The methods associated with evidence-based measurement are extensive, and may appear daunting to students, and possibly even to teachers who do not specialize in them. However, even if students do not develop a fluent knowledge of how to execute these methods, a basic comprehension of why they are used, the rationale behind them and which ones to look for is important as scaffolding for competently and consciously engaging with instruments in a critical rather than naive way. Before outlining these, it is worth drawing a distinction between classical test theory (CTT) and item response theory (IRT) which is sometimes also referred to as modern latent trait theory. Classical test theory as the name would suggest has a longer pedigree than IRT, and is based on true score theory which most fundamentally states that any observed score is comprised of the true score as well as some error (Nunnally and Bernstein, 1994). We are always interested in the true score, but the true score is never observed perfectly in practice, because it is always associated with some degree of error. The associated methods are founded on the relationship between these three aspects, and in CTT items are assumed to be parallel, or put another way, of equal difficulty. IRT, on the other hand, has as its premise precisely that items are not of equal difficulty, or at least cannot be presumed to be, and that it is a positive attribute if they are not. In IRT, the focus is at the item level rather than the overall test level, and the response to an item is a function of both an item and person parameter (Hambleton, 1989; Rogers et al., 1991). The approach heralds some major advances in areas such as computerized adaptive testing whereby tests can be made shorter. IRT is more sophisticated and more difficult to grasp, and given that the focus of this paper is on students who need some understanding of measurement, but whose field specialization is not such, or even somewhat associated with such, I will restrict my coverage of methods to the core methods associated with CTT which would include indexes of reliability and factor analysis. Furthermore, the majority of psychometric studies students will encounter with respect to self-report inventories will still relate to CTT rather than IRT. While IRT continues to gain an expanding presence in the literature, methods founded on CTT are still very prevalent.

Turning then to core methods, associated with CTT, which should form the basic toolbox, so to speak, of a student's critical engagement with the use of self-report inventories for mental measurement, there are two important categories here; and these are the methods students are most likely to encounter. The first is factor analysis which can be separated into two further subcategories, namely, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The second is indexes of reliability, of which there are many. Factor analysis is more protracted and complex in its execution than any particular index of reliability, and is typically associated with the development (EFA) of a measurement model for a self-report instrument, and then the confirmation (CFA) of a measurement model for the same tool (Kline, 1994; Kline, 2005; Henson et al., 2004; Henson and Roberts, 2006). Indexes of reliability are typically associated with the routine use of an instrument which should, in terms of, and assuming best practices, have already undergone analysis using factor analysis. It could be said that indexes of reliability, in general, presume a measurement model rather than demonstrate one, and so one could say that there is something of a natural order here. Factor analyses occur first and are more foundational, and indexes of reliability can occur thereafter as the instrument is deployed for general use, and for research purposes which need to presume good measurement but which may not have the analytical space to conduct factor analysis.

The division of factor analysis into EFA and CFA and their appropriate use is of critical importance. Essentially, both are forms of data reduction, in the sense that we are interested in the extent to which a group of items can be reduced, through the variance they share, onto a common and underlying factor, and with this factor representing the latent or construct. A factor, which is fundamentally an expression of shared variance among the items which comprise it, is therefore not classed as an observable (the individual items are), but something latent which is said to cause the scores derived on the items. Because this cause is common to a group of items, this group of items should, in principle, present with a lot of shared variance, and be reducible to the same factor. This is the underlying principle of data reduction which applies to both EFA and CFA, and it assists with mapping items onto constructs or latents and, therefore, in articulating a measurement model. The measurement model, in turn, expresses which items measure which constructs in any given self-report instrument, and therefore also expresses what the scoring regime for items comprising the instrument should be.

There is an important difference between EFA and CFA however, and this is often not well understood, even in some cases by developers of self-report instruments. In EFA, one is concerned with letting the data express itself, so to speak; and in this sense we arrive at models which are led by the data, although these models are not independent of certain executional decisions which impact how the data expresses itself. It is for this reason that EFA is often referred to as producing a posteriori models; i.e. models which are after the fact of the data, and therefore directed by the data. It is also for this reason that EFA is primarily a development tool, because it can be used to explore the dimensionality of data which presents when a range of candidate items have been developed for an initial attempt at representing some or other construct. In CFA, and on the contrary, we approach the data with a model in mind, and directly test the fit of this model against the dimensionality of scores in the data. In a CFA-based test of model fit against a set of data, the model is often referred to as being a priori because the model precedes the data. In other words, we have a hypothesized measurement model in mind for an instrument, with this model being led by theorizing or prior development using EFA, and then we go out and collect a new dataset and test the plausibility of the model in that dataset. Put another way, we have a conception about which items measure which constructs, and we want to directly test this conception to confirm whether it is fundamentally correct, and whether there is positive evidence for it. It is important to note that if a dataset were used to arrive at a measurement model, a test of the model using CFA in the same dataset would no longer constitute an a priori test, because the model emerged out of the same data, or is a posteriori to that data. The test would only be a priori if the model were tested in a new sample collected from the same population. Data could be collected from other populations, but this would typically be to establish whether a model we know to fit in one population also fits in another population, therefore extending the populations in which the instrument is known to measure satisfactorily. It is a general principle that evidence for a model in one population does not count as evidence in another.

There are some differences between EFA and CFA associated with the different rationale which underpins them. In a model emerging from an EFA, all items will have been permitted to operationalize, or indicate, all factors (with the number of factors, representing latents, having been determined by some prior executional decisions). This assists with giving us an idea of how much a given item is caused by the respective factors in the model. Ideally, an item should be caused by one factor (representing the latent), and one factor only, permitting the claim that the item is an exclusive measure of the latent. The claim to exclusive measurement is critical to having a measurement model, and associated evidence-based scoring regime, because if any given set of items is caused by more than one latent, when they are added up, or averaged, for a composite score, we do not know what the composite score means. Put another way, the interpretation of the composite score is essentially indeterminate, because there is more than one latent causing it. In practice, however, exclusive indication of a latent by an item is never the case in human measurement, and the fact that an item can indicate any of the factors in the model, gives us the opportunity to determine the extent to which this is the case; i.e. the extent to which the ideal has been met in practice. What this means is that items are essentially allowed to "cross-load" in an EFA. In CFA, the principle is precisely the opposite. Items are specified in a hypothesized model to indicate one factor only, and are not permitted to cross-load. So a model might comprise five items hypothesized to indicate Factor A, and only Factor A, and five items hypothesized to indicate Factor B, and only Factor B. This entire model of two factors and 10 items is then tested directly against a set of data to determine whether it fits the data, and as with other branches of statistics, the test is conducted as a function of probability (chi-square test); in this case, the probability that the measurement model is commensurate with the actual dimensional properties of the data. However, and somewhat particular to CFA, additional indexes are also used to determine fit because the chi-square test has a tendency to over-reject models, even if the departure of the model from the dimensionality of scores is fairly trivial, and especially so when large datasets are being used (Kline, 2005; Byrne, 2001; Byrne, 2005).

The above account of the distinction between EFA and CFA is necessarily brief given the scope and purpose of this paper. However, the important issue is what needs to be abstracted from this from a pedagogical point of view, and with respect to what students need to understand about what these methods are, their place in the literature, and their place in evidence-based measurement using self-report instruments. The central issue, both from the critical perspective of evaluating studies which have used an instrument for making research inferences, and from the perspective of adopting an instrument for research for the student's own research purposes, is that there should be studies indicating that the instrument in question has been developed with an explicit measurement model in mind. Furthermore, this measurement model should have been confirmed using CFA as the method and in the population for which it is intended to be used. Of course, it may, or may not, be the case that EFA was used to develop the instrument, and if it was, this is commendable, but in the final analysis, the measurement model requires confirmation. In the case that an EFA has been conducted but not a CFA, and this is frequently the case because CFA is more difficult and dedicated software is required, then the results of the EFA would be illuminating, but would not count as confirmation of the measurement model. Also, if there are a series of EFAs which have been conducted on different samples taken using the same instrument, the student should expect these to vary in terms of the measurement model they suggest for the instrument, and this is because EFAs tend to be sample dependent (Henson et al., 2004; Henson and Roberts, 2006), and also involve a sequence of decisions which are a question of judgment, and which tend to vary. These may include decisions like interpreting criteria on how many factors to extract in the analysis, and the threshold used to decide if a coefficient points to a particular item as being significantly caused by a factor. It is this instability of derived solutions across studies in EFA which makes CFA even more necessary. With CFA, one tests a presumed model directly against the data and the question of whether the model fits the data or not is relatively determinate. With EFA, the combination of the data and a series of decisions made by the researcher leads to solutions which can vary, and while these solutions are not arbitrary (because one presumes the researcher to have made good ones based on the evidence in front of him/her), they do tend to vary across studies.

Immediately above, it was stated that there should be CFAs which have been conducted in the population for which the instrument is intended, and this qualifier with respect to the population is very important within CTT, and reiterates a point already made above. If an instrument has been used in one population, with confirmatory evidence for its measurement model in that population, it cannot be presumed that the measurement model would fit if the instrument were used in another population (Hambleton, 2005). Here the issue of population specification is extremely important, and the issue is so extensive that it cannot be dealt with in this paper, because there are many parameters for specifying a population. However, to take one clear and very simple example, if one takes an instrument designed for the North American population (which will almost always be in English) and then translate it into Japanese for use in the Japanese population, confirmation of the model in the original population of North America does not suffice when the translated version is used in Japan; the results of the CFA in the North American population are not transferable, so to speak, to the Japanese population, and do not count as evidence for the model in the Japanese population. A new and separate CFA would have to be conducted in the Japanese population to confirm that the measurement model is appropriate and fits for the Japanese-language version when used in Japan.

From a pedagogical point of view, instruction in the place and role of factor-analytic methods in the development of good self-report instruments is vital. It assists students with being able to critically engage with the question of whether an instrument they have encountered, or they have decided to use for their own research purposes, has actually undergone serious development and is likely merit-worthy. This is particularly important precisely for the reason that so many instruments are not merit-worthy, and results premised upon the data they generate is founded on measurement for which there is little, or no, evidence. Beyond the initial development of instruments, however, there is also the issue of the routine use of instruments even when, and after, evidence for the structural validity of scores has been demonstrated through factor analytic methods. Such routine use of instruments is often attended by the reporting of results for indexes of reliability. The value and use of these indexes, and what they are able, or not able, to demonstrate, should also be critically understood. The most commonly encountered index of reliability is Cronbach’s alpha, and in fact its presence in the literature is so pervasive that Sijtsma (2009) has pointed to the original article presenting the index (Cronbach, 1951) as having been cited more often than the seminal article reporting the identification of the double helix as the structure of the DNA molecule. Unfortunately, the pervasiveness of the index is somewhat proportional to the lack of awareness concerning its limitations. While the word “limitations” has a necessarily negative connotation to it, it should also be said that from a pedagogical perspective, the analytical limitations of something, like the Cronbach’s alpha reliability index, also have the very positive characteristic of being the potential starting point for developing greater critical awareness. In the case of Cronbach’s alpha, the primary limitation is the incapacity of the index to speak to the unidimensionality of a measured construct (i.e. the notion that the items producing the scores for a particular construct, measure only that construct and no other) and this has been well-covered in the literature (Green et al., 1977; Cortina, 1993). The index cannot, therefore, stand in for CFA which is able to demonstrate unidimensionality, and awareness of this limitation, often unseen by people who uncritically take the pervasiveness of the index as an endorsement of it as some sort of panacea, serves as a very useful pedagogical starting point for greater critical awareness among students engaging with the index. Another, limitation of alpha is that it is positively biased by the number of items comprising a measured construct (Green et al., 1977; Cortina, 1993), and this should be considered when interpreting the value returned for it on any particular set of items comprising a measured construct, and of course also presents as a pedagogical starting point for a critical awareness of using alpha. Alpha is the most common index encountered in the literature and presumes interval scales. There are other indexes for interval scales, and other indexes for scales which are not interval. Students can become familiarized with these and their respective strengths and limitations, without necessarily having a deep understanding of the mathematics which underpins them.

Issues Hidden behind Statistical Method in a Critical Approach to Self-Report Instruments

The section above, though necessarily brief, covers some of the essential statistical methods involved in a critical approach to self-report instruments, and methods which students should at least be familiar with, if not fluent in. However, it is worth elaborating that statistical methods should not stand in for a thoughtful approach to such instruments, even when the results of such methods are superficially positive. A fully developed pedagogy with respect to preparing students for critical engagement with these kinds of instruments should not be represented as being circumscribed by statistical training alone. It is quite possible to derive good results from analyses involving factor analysis (of either type), and yet still have scales which require further critical analysis, and for which such analysis, if conducted, might indicate serious problems. Students need to be sensitized to this so that their critical posture exceeds a mere subservience, or intellectual deference, to statistical methodology. One example of such a situation would occur when the items comprising a factor (which of course represents the underlying latent of interest) are near repetitions of each other but with slightly different wording. The problem with this is that each one of the items is supposed to represent a separate and different operationalization of the underlying construct of interest, and the probably good results for the factor derived in a statistical analysis (whether through factor analysis or indexes of reliability), would in fact be artificially derived. This would be because the same operationalization is being repeated by multiple items, and it would therefore not be surprising that these items shared a lot of variance. This artificial derivation of good results would also, therefore, come at the expense of the full and proper representation of the underlying construct through diverse operational expression. In such a situation, the individual items have essentially become identical expressions of the underlying construct, thus essentially becoming redundant, with this being disguised only by the slightly different wording in each case.

Beyond the issue of item redundancy, and even when the claimed operational expression of the construct is diverse, there is still the issue of item interpretation and the reasoned account of an item's place in a construct (i.e. how the behavior represented in the item actually does express the construct of interest); and this is not to mention the issue of the extent to which all the items appearing for a construct actually do represent it. This is not an issue for discovery by statistical method, but rather by reasoned engagement with the content and meaning of the items themselves. This is best illustrated by concrete example. The five items below are claimed to represent the preference for visual learning construct in the Personal Learning Styles Preference Questionnaire (PLSPQ; Reid, 1987; Reid, 1990; Reid, 1984):

6. I learn better by reading what the teacher writes on the chalkboard.
10. When I read instructions, I remember them better.
12. I understand better when I read instructions.
24. I learn better by reading than by listening to someone.
29. I learn more by reading textbooks than by listening to lectures.

(Note: The numbers in front of these items indicate their ordered position in the 30-item questionnaire which also claims to measure five other constructs related to perceptual learning style preference)

The preference for visual learning construct in the PLSPQ is entirely represented by these five items, which appear on a subscale labeled Visual. A student who has not had the benefit of some form of instruction leading to a more critical engagement with self-report instruments might be ready to take the label “Visual” at face value. However, more careful consideration, and without the assistance of statistical methods at all, should lead to important critical questions with respect to what is exactly represented by these five items. Notably, every one of them places “reading” as the central behavior in the item, and while reading may involve vision because we have to see text to read it, it does not fully circumscribe the construct of visual learning. People also learn visually through schematics, charts, color-coding of categories, pictures, visual mnemonics and the list goes on; with many of these examples involving no reading or engagement with text. Furthermore, reading, in addition to not fully circumscribing the construct of visual learning, also circumscribes modes of perception which are not visual. Most people describe some form of inner-voice (or silent speech) when they read, and although not heard out loud by third parties, and although not a direct form of oral production and auditory perception, it is very arguable the internal expression of such (Vygotsky, 1978; Vygotsky, 1986) and phenomenologically reaches beyond visual learning. Besides all this, there is the mere fact that most people, for most of the time, experience reading individually; i.e. as something they do for themselves to extract and comprehend information. In this sense, reading has significant overlap with a private learning construct, and one may even be tempted to relabel this construct of Visual in the PLSPQ as the Private Learning Construct. This example clearly illustrates how the inference made about the label for a subscale on the basis of the items’ content is an inference subject to critical interpretation; and this critical interpretation involves reasoning which is not statistical and which is quite feasible for students with no grounding in statistics.

Conclusion

Many humanities students, whether undergraduate or graduate, are not pursuing fields where knowledge of formal, statistical measurement is an explicit requirement or emphasized, although this is bound to change as such forms of measurement creep into disciplines long pursued under an entirely scholarly approach. This shift to a more prominent role for formal measurement in disciplines not traditionally familiar with them, progresses, therefore, as something of a methodological development, and curricular responses will inevitably follow with time. However, the object of knowledge for the humanities, rather than its method, is naturally the domain of the humane, and this domain is increasingly permeated with the human experience of being measured formally and explicitly. In this sense alone, understanding formal measurement becomes a core requirement for understanding the modern human condition. This is for the very reason that it is increasingly the personal experience of so many people who live and compete in modern industrial and information societies and whose fates depend on it to one degree or another. One precondition for understanding this lived experience is, at least, having some grasp of how the formal measurement of people, and their minds and behavior, actually works. Given the ubiquity of measurement to the modern lived experience and the enormous variety of contexts in which it is encountered, laying out a full curricular and pedagogical response for the humanities is a considerable task which certainly exceeds the capacity of a single paper. However, the issue pleads for attention, and some kind of curricular response, in humanities faculties aligned predominantly, sometimes almost exclusively, with a traditional scholarly approach to disciplinary activities. Also, given the scope and magnitude of the issue, such responses should begin at the undergraduate level and continue through to the postgraduate level.

Overall, it befits programs within the humanities to promote an increased awareness of, and critical engagement with, formal measurement. The selection of self-report instruments for discussion in this paper, besides covering one of the most frequently encountered forms of measurement tool, presents pointers as to how, at least, some of this critical engagement might be achieved. The conundrum that often presents with a curricular response to teaching aspects of formal measurement is that the statistical methods employed are difficult and time consuming to learn. This, unfortunately, can result in neglect of the problem for the perceived magnitude of what it would take to engage with it. However, and in fact, a conceptual understanding of how such measurement works in principle can be achieved with relatively little opportunity cost in terms of time allocated out of an already challenging curricular schedule. In the case of self-report instruments, which occupy a significant quotient of the forms of measurement students will encounter both inside the academy and outside in the world of work and life, much can be accomplished with a relatively skeletal pedagogical and curricular approach. A foundation in the principles of latents and observables can underpin a further familiarization with the associated statistical methods, i.e. methods related to data reduction including EFA and CFA, even if knowledge of these methods does not proceed to a full executional knowledge. This association of principles and methods can then be elaborated under the general rationale of explaining how the methods facilitate an interrogation of the correspondence between latents and observables in their expression of a measurement model. This alone, even without a deep understanding of the mathematical procedures and types of statistical estimation behind the reduction, puts students on notice that there needs to be evidence for measurement, before the data from measurements are used as evidence for anything else. This limited foundation given to students may not always be enough for them to critique the poor execution of statistical methods which would require a deeper understanding, but it would certainly be enough to pick out those instruments which have entered the research literature or been deployed in practice in the absence of the use of such methods at all, and which is often the case.

Endnote:

(1) It is worth mentioning that the example here could be conceptualized differently whereby extraversion and introversion are opposing poles of a single bipolar construct. However, for the purposes of illustration extraversion and introversion are conceptualized here as two separate constructs.

REFERENCES

Byrne BM. (2001) *Structural equation modeling with AMOS*, London: Lawrence Erlbaum Associates.

Byrne BM. (2005) Factor analytic models: Viewing the structure of an assessment instrument from three perspectives. *Journal of Personality Assessment* 85: 17-32.

Cortina JM. (1993) What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology* 78: 98-104.

Cronbach LJ. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika* 16: 297-334.

Fitzpatrick K. (2012) The humanities done digitally. In: Gold MK (ed) *Debates in the digital humanities*. Minneapolis: University of Minnesota Press, 12-15.

Goldman, B.A. and Mitchell, D.F. (2008). Directory of Unpublished Experimental Mental Measures. Washington, DC: American Psychological Association.

Goodson S. (2012) If you're not paying for it, you become the product. *Forbes*. Retrieved from https://www.forbes.com/sites/marketshare/2012/03/05/if-youre-not-paying-for-it-you-become-the-product/#4fcdb08c5d6e

Green SB, Lissitz RW and Mulaik SA. (1977) Limitations of Coefficient Alpha as an Index of Test Unidimensionality. *Educational and Psychological Measurement* 37: 827-838.

Hambleton RK. (1989) Principles and selected applications of Item Response Theory. In: Linn RL (ed) *Educational measurement*. 3rd ed. New York: Macmillan, 147-200.

Hambleton RK. (2005) Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In: Hambleton RK, Merenda PF and Spielberger CD (eds) *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum, 3-38.

Henson RK, Capraro RM and Capraro MM. (2004) Reporting practices and use of exploratory factor analyses in educational research journals: Errors and explanation. *Research in the Schools* 11: 61-72.

Henson RK and Roberts JK. (2006) Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement* 66: 393-416.

Hockey S. (2004) The history of humanities computing. In: Schrelbman S, Siemens R and Unsworth J (eds) *A companion to Digital Humanities*. Oxford: Blackwell, 3-19.

Kline P. (1994) *An easy guide to factor analysis*, London: Routledge.

Kline RB. (2005) *Principles and practice of structural equation modeling*, New York: The Guilford Press.

Marsh HW. (1984) Students' evaluations of university teaching: Dimensionality, reliability, validity, potential baises, and utility. *Journal of Educational Psychology* 76: 707-754.

Marsh HW. (2007) Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In: Perry RP and Smart JC (eds) *The scholarship of teaching and learning in higher education: An evidence-based perspective*. Dordrecht: Springer, 319-383.

Moretti F. (2007) *Graphs, maps, trees: Abstract models for a literary history*, London: Verso.

Nunnally JC and Bernstein IH. (1994) *Psychometric theory*, New York: McGraw-Hill.

Reid JM. (1984) Perceptual Learning Style Preference Questionnaire. Copyrighted by Reid. Available through Joy Reid, Department of English, University of Wyoming, Laramie, WY 82070.

Reid JM. (1987) The learning style preferences of ESL students. *TESOL Quarterly* 21: 87-109.

Reid JM. (1990) The dirty laundry of ESL survey research. *TESOL Quarterly* 24: 323-338.

Rogers HJ, Swaminathan H and Hambleton RK. (1991) *Fundamentals of item response theory*, Newbury Park: Sage.

Shevlin M, Banyard P, Davies M, et al. (2000) The validity of student evaluation of teaching in higher education: love me, love my lectures? *Assessment and Evaluation in Higher Education* 25: 397-405.

Sijtsma K. (2009) On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* 74: 107-120.

Solon O. (2011) You are Facebook's product, not customer. *Wired*. Retrieved from http://www.wired.co.uk/article/doug-rushkoff-hello-etsy

Spooren P, Brockx B and Mortelmans D. (2013) On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research* 83: 598-642.

Vygotsky LS. (1978) *Mind in society: The development of higher psychological processes*, Cambridge, MA: Harvard University Press.

Vygotsky LS. (1986) *Thought and language*, Cambridge, MA: MIT Press.

Wachtel HK. (1998) Student evaluation of college teaching effectiveness: a brief review. *Assessment and Evaluation in Higher Education* 23: 191-211.

Wolf A. (2002) The growth of psychometric testing. *British Educational Research Association Annual Conference*. University of Exeter, Retrieved from http://www.leeds.ac.uk/educol/documents/00002458.htm.