Journal of Social Sciences 3 (3): 106-116, 2007
ISSN 1549-3652

Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes

[1]James Carifio and [2]Rocco J. Perla
[1]University of Massachusetts Lowell, One University Ave, Lowell, MA 01854
[2]HealthAlliance Hospital,60 Hospital Road, Leominster, MA 01453

Abstract: A recent article by Jamieson in Medical Education outlined some of the (alleged) abuses of "Likert scales" with suggestions about how researchers can overcome some of the (alleged) methodological pitfalls and limitations[1]. However, many of the ideas advanced in the Jamison article, as well as a great many of articles it cited, and similar recent articles in medical, health, psychology, and educational journals and books, are themselves common misunderstandings, misconceptions, conceptual errors, persistent myths and "urban legends" about "Likert scales" and their characteristics and qualities that have been propagated and perpetuated across six decades, for a variety of different reasons. This article identifies, analyses and traces many of these aforementioned problems and presents the arguments, counter arguments and empirical evidence that show these many persistent claims and myths about "Likert scales" to be factually incorrect and untrue. Many studies have shown that Likert Scales (as opposed to single Likert response format items) produce interval data and that the F-test is very robust to violations of the interval data assumption and moderate skewing and may be used to analyze "Likert data" (even if it is ordinal), but not on an item-by-item "shotgun" basis, which is simply a current research and analysis practice that must stop. After sixty years, it is more than time to dispel these particular research myths and urban legends as well as the various damage and problems they cause, and put them to bed and out of their misery once and for all.

Keywords: formats, Likert, measurement, psychological, scales

INTRODUCTION

In the process of reviewing literature related to assessments in medical and health education, we came across a recent article by Jamieson in Medical Education that attempts to outline some of the (alleged) abuses of Likert scales with suggestions of how researchers can overcome some of these methodological pitfalls and limitations[1]. However, many of the ideas advanced in the Jamieson article relative to Likert "scales," as well as a great many of articles it cited[2-6], are themselves common misunderstandings, misconceptions, conceptual errors, myths and "urban legends" about Likert scales and their characteristics and qualities that have been propagated and perpetuated across decades, for a variety of different reasons, including a lack of first hand familiarity and understanding of primary sources (i.e., Likert's actual writings), and various and definitive primary empirical studies done by Likert and others (see below). In this respect, Jamieson is no different from the dozens of sources over a twenty year period she cites in her article about "Likert scales." Further, this problem is not just confined to the field of medicine and medical education, as the majority of the articles that are the source and propagators of many of the most important errors and misunderstandings currently extant concerning "Likert scales," are from psychology, education and the field of psychometrics in the fifties and early sixties[7-13]. These "root of current urban legend" articles, moreover, are additionally more than just "historical curiosities" to anyone who has actually read Likert in the original or constructed and empirically developed a "Likert scale" according to his theoretical model and writings[14,15]. This article, therefore, addresses this important problem and a number of persistent misconceptions, misunderstandings, and factual and empirical errors, myths and untruths about Likert scales and their characteristics and properties with the hope of helping

Corresponding Author: James Carifio, University of Massachusetts Lowell, USA

researchers and practitioners understand the various factors, complexities, specifications and sophisticated nuances that must be considered whenever any given measurement scale (or response format) is used, developed, or analyzed. Further, one of the central points in this article for medical and allied health educators and researchers (as well as those in other fields) is that the same level of skill, ability, theory, and rigor that goes into all scientific and biomedical measurements is also required in educational measurements of all types and kinds for the serious educational scholar and researcher, as the principles of scientific measurement are the principles of scientific measurement (and the "heart of science") in virtually all domains[16].

DISCUSSION

Scales versus Response Formats: One of the primary confusions in the Jamieson article centers on the use of the word scale (versus **response format**). Clearly, the author, similar to a large number of the sources she cites, is referring to a **response format** as opposed to a (measurement) **scale** (see below) in her discussion, yet no distinction whatsoever is made between the two, as if such a distinction is either unimportant or does not exist, both of which could not be further from measurement theory or the truth, or Likert's actual and original writings on these matters[14]. This particular point is so central to accurately understanding a Likert **scale (**and other scales and psychometric principles as well) that it serves as the bedrock and the conceptual, theoretical and empirical baseline from which to address and discuss a number of key misunderstandings, urban legends and research myths.

Distinguishing between a **scale** and a **response format** is not always easy to do, or straight forward, because it first requires some linguistical analysis and close attention to word and term meanings, and the contexts in which the word or term are used. This particular point is true and important and needs to be understood as both "measurement" and "statistics" are areas of poor, careless, ambiguous, confusing, and misleading language usage, as well as areas of profuse and unthoughtful usage of a wide variety of professional slang. Of particular importance is the fact that the word and term **scale** and **response format** in the domains of measurement and statistics is like the word **"interval"** in these same two domains, which has several different specific meanings; namely, interval scale, data interval (obviously different from scale), confidence interval, and so on, as "interval" is a generic idea and concept that is used, defined and particularized in many important different ways in both of these domains. The key here is that the word "interval" has a qualifying term (adjective) in each of these instances. The problem, however, is that the absence (or implied presence) of the appropriate qualifying terms in a given content can create many confusions, misunderstanding, and errors of various kinds. There are many such terms, words and concepts in educational, psychological, and sociological measurement. Further, linguistic sloppiness or carelessness and **slang (or "techie") usage** of such words and terms by people doing work in these domains (and most particularly the alleged "specialists") is one of the major sources and causes of difficulties, which leads to multiple confusions, misunderstandings, errors, myths and urban legends, and particularly so for novices, or someone new to the particular sub-area in this field (a prime example of these points is the term "logit regression").

To clarify this problem further and elucidate its many facets, consider the following linguistic/conceptual problem. The (fictitious) 20 item Box personality test (which is a **scale**) has a binary **response format** (what would carelessly and inaccurately be called a "scale" by the majority of professionals today, which as will be seen below, has nothing to do with it being binary). The sloppy and incorrect language (and thus meaning and conceptualization) that one typically encounters relative to this example and statement (and by measurement and psychometric professionals who are often the worst of the offenders) is "the Box personality scale has a binary response scale," where the meaning of the term binary response scale in the statement is **connotatively referring to a particular data type (i.e., a nominal [data] scale).** This impoverished "techie slang speak" (TSS) is a careless, colloquial, (and connotative) usage of the word **scale** for (and to mean) data type. So we now have 3 different usages and meaning of the word "scale" in one sentence; namely, the (real) measurement scale (the Box test or instrument), the scalar properties of the response format (or lack thereof), and the data type (often also confusing called the "measurement scale" of the data). Also, it should be noted that (truly) nominal data is held not to be a scale at all because it has no underlying continuum, so the errors and carelessness is further compounded if we do not define the binary categories of the "scale" (i.e. response format). If the binary categories are "agree" and "disagree" as opposed to "yes and no" or "true and false," then we have a severely truncated ordinal response format (and data type) as opposed to a

nominal response format (and data type), and there is an underlying continuum even though the response format is binary. Note well and clearly, please, that the adjective "ordinal" in the previous sentence is a scalar property of the item response format (and not the 20 item instrument, which is the real scale) and that this ordinal characteristic is actually something more than just the property of the item response format alone, as will be seen more specifically below.

Language, at every level, therefore, is not only critically important, but it should also be clearly noted that just about every intelligence and achievement test reduces the item response format used to binary form (correct and incorrect) and both are considered to be and treated statistically as **interval scales,** which starkly contradicts logically and empirically most of what Jamieson and the many "authorities" she cites have to say about response formats, scale types, and the do's and don'ts of statistically analyzing and interpreting them. All of these points also emphasize and illustrate that there are critical conceptual and operational differences between a response format (information capture protocol or device) and a response format scoring (meaningful coding) procedure (or protocol) that is used to transform the information captured into an element or unit of an interpretive system, and, hopefully, theory of some kind. Some response formats **fuse** these two item components (the capture and scoring/coding of the information) and have each subject (i.e., respondent) do their own "coding" of the (covert) information the subject is processing/experiencing in real time (thus reducing researcher burden and costs), while other response formats **do not fuse** the capture and coding component of the generic 3 component **standard "item" model** (i.e., the stem [stimulus or question etc], response capture procedure/device, and the response transformation into meaning units [coding or scoring etc] item components). **So "scalar" properties at the item level tend to be the properties of the data transformation [coding/scoring] component** of the standard item model rather than the response format or stem components, although scalar properties may be built into these components also for a variety of different reasons, some practical and some theoretical. So in this model of an item, an item in a patient examination protocol or **scale** would be: "open please" (stem), insert thermometer, wait appropriate amount of time, visually observe digital value (information capture), say "you're not sick today my boy, back to work with you" (coded and interpreted response). It should be noted that the third component (the coding and interpreting) implies a rule of some kind and a norm of some kind (i.e., a fourth component) that is either embedded or explicit and detached (fused or not fused), but this fourth component will be discussed later. Given this "standard item" model, it would seem fair and reasonable to ask, "Why are "Likert scales" so widely misunderstood, and why are the many contradictions in the erroneous view, myths and urban legends identified to this point in this article so glaring?" The answers lie in "levels and units" of analysis and emergent properties of collections of items.

Atoms, Molecules and Scales: A group of questions that have nominal response formats, particularly if the number of questions in the group is large enough, can be a scale (such as the fictitious Box personality test or **scale** referred to above), and in fact an **interval scale** (data type), and even a **ratio scale** (if, for the sake of argument and making a point, the response format anchors were "never" and "always"), if the group of questions (items) has the necessary logical and empirical properties (see below). So a scale, in this meaning and sense of the term, is an **emergent property (**i.e., molecule) of the group of items (atoms) and the **properties** of the items (both logical and empirical) **that connect them together into a whole** (i.e., molecule). Such a scale (i.e., molecule) is a **measurement scale,** which has a more complex meaning than the individual items (atoms) that comprise it, or the **different parts of these items** (e.g., stems, responding formats, and scoring/coding procedures), and one does not really have what is truly meant by the word and term **scale** in measurement and psychometrics until one has a minimum group of such observations (i.e., 8 items at a minimum usually), as it is this measurement scale on which one obtains the required reliability and validity needed to be able to use, analyze and interpret the data collected.

So, there is a world of conceptual and empirical differences between an **item responding format** (and what data type or "scale" the responding format is if it is a fused one), and a **measurement scale** (and what data types the **derived indices** from such a scale are). The language, qualifiers, contexts and precisions of expression used are, therefore, not trivial or mere semantics, as demonstrated in detail in this and the preceding paragraphs, but they are key, and key to appropriate understanding, conceptualization and communication, and to the avoidance of various Ryle-like category mistake and classification errors and their associated misconceptions and misunderstandings. As

Likert continually pointed out, there is a vast difference between a Likert **responding format** and a Likert **scale**, and this key fact and difference got "lost" in the literature in the 1950's and confused by research specialists and psychometricians who were very linguistically and conceptually careless, and who did not bother to read Likert for a number of different reasons. One aspect of the problem, then, is that the language and conceptual "house cleaning" and "house keeping" that should be routine in the measurement and psychometric literature is not present now, and has not been very good in the past, and these facts have concerned a number of scholars recently[17]. These two areas (measurement and psychometrics), therefore, are strewn with various often disguised difficulties and pitfalls for those unacquainted with the history of these two disciplines somewhat like the emergence in quiet villages of dangerous fields long after a war has occurred and everything now is seemingly tranquil until someone unwittingly mis-steps and sets off an old "buried bomb."

The key points made about measurement scales in the paragraph above is that **scale items are not autonomous and independent** (i.e., the behaviorist and blind empiricist view), but rather they are **a structured and reasoned whole** (particularly in neo-positivist/cognitive models), which also meet certain empirical criteria as well as logical and content criteria. There are five (5) basic and widely agreed-upon kinds of validity in the psychometric literature[18], but the five may be conceptually reduced to logical/semantic (content and face validity) and empirical (concurrent, predictive and construct) types of validities, with the empirical validities being confirmatory of the logical validities (i.e., concept/theory and observed facts/agreements). The major problem with blindly empirical and logical positivist models and views of measurement and these concepts, such as those suggested in the article in question that is being critiqued here as well as a vast number of other sources[19-21], is that the logical requirements and components of scales, items and responding (and automated self-coding) formats are almost totally disregarded or ignored or treated as a fuzzy jumble. The simple fact is that they cannot be treated this way or ignored—ever; that is more than "reductionism with a vengeance" and "over-simplification."

Given the points made above, examine the sentence: "The (fictitious) Box Personality test (scale) has a Likert response format," or even more germane to this discussion: "The Likert Attitude Towards Measurement Scale has (uses) a Likert response format." This second sentence is more demonstrative of the communication and conceptual difficulties and various language traps faced in educational and psychological measurement because (1) it is recursive and self-referential (particularly if not appropriately "decoded" and 'rewritten"), and (2) there is **absolutely no requirement** for the items of a Likert Attitude Scale **to have a Likert response format.** This particular point is an extremely critical point to a vast array of erroneous, mythical, illogical, and jumbled statements and claims that are made about "Likert scales" in the "literature." In fact several different types of response formats could be used if they have certain characteristics (e.g., a 100 millimeter line or continuum with semantic anchors on each end).

Studies done by Carifio[22,23] showed that using a 100 milliliter line with 2 to 7 anchor points as the **responding format** to attitude statements or semantic differential stem phrases produced data that was **empirically** linear and interval in character (as both properties may be empirically tested for any scale or dataset) at the subscale and full scale level. Further, the data from this response format correlated to responses made to the same questions using a 5 to 7 point "Likert" response format at r=+.92 (N=457). This high level of correlation means that the data obtained from the two response formats and the properties of the data obtained from the two response formats are highly isomorphic. These data were collected from subjects who ranged from the eighth grade level to adults. The findings from these studies were supported and extended by other researchers[25-27]. The Vickers study in particular is an excellent exemplar of how "Likert scale" data can very closely approximate **ratio** scale data under a particular set of conditions[27]. These very basic empirical facts (and studies) contradict several of the (armchair and "logical") claims made by Jamieson, and a large number of the authorities she cites (including many venerable experts in the psychological, measurement, psychometric and educational literature), about the conceptual and empirical nature of "Likert scales" (i.e., self-coding response formats), but a fuller interpretation of these basic empirical facts in combinations with others will be postponed for now until other points have been made.

Macro and Micro Measurement Levels: What we see in the various confusions in this literature is that the word "scale" as subject and the word scale as "predicate (object)," as well as the word scale as adjective and as a process (i.e., scaling of usually coded responses), all have different meanings. We also see the word scale at

the **macro level** (i.e., a collection of purposefully constructed items according to an a priori blueprint) and at the **micro level** (i.e., the manner in which one responds to or provides information, self-coded or not, about each item in the collection of items, which is the response format and certain characteristics of the response format) are and mean very different things. For example, is there an underlying continuum and what is the nature of this continuum at the item level and then again across the collection of items at the (macro) scale level? Alleged measurement and statistical experts and commentators often write about "Likert scales" when they mean Likert (self-coding) response formats, and they perpetuate various urban legends, inaccuracies and untruths about each. They fail to make this critical distinction or understand the critical and logical as well as **empirical differences** between the two. They also fail to distinguish, understand or to be extremely clear and precise about at exactly what level (macro or micro) and for what model components the various scalar characteristics they are talking about are the characteristics they claim rather than conceptual errors and category mistakes they are making, otherwise they might clearly understand how ordinal item response formats can and usually do produce scales that are empirically interval level scales. Many of the nuances and facts of these central problems, and the misunderstandings these commentators and authorities create and perpetuate are discussed in more detail below.

One Swallow: The Likert response format is only a problem, as opposed to various commentators claims to the contrary, **if one analyzes each individual item on a scale or questionnaire separately,** which one should not ever do because of the family wise error rates of repeated statistical testing (never mind the "blind shotgun empiricism" research approach), and the fact that **a single item is not a scale in the sense of a measurement scale** (i.e., "one swallow a summer do not make"). This particular analysis practice is one of the very poorest of research practices, unless one is doing item analysis or very formative and exploratory analysis of one's measurement scale and research questions, which is not the point where the analysis or research process stops. One way to understand this point is to answer the following questions: "What do you think of a **one-item IQ test** (that was scored as either right or wrong)**,** and what would you say (namely, what would your arguments be) if your rival got the item right and you got it wrong?" How many one item IQ tests are there in Burros' Mental Measurement Yearbook? How many of those tests are validated, and how do you assess a theory of multiple intelligence with a one item IQ test? So, it should be reasonably clear that in almost all situations and circumstances "**One item a scale doth not make**." What then makes someone (or you) think that it is completely appropriate to analyze your questionnaire or Likert scale item by item and then to present this "unorganized laundry list" and "fuzzy jumble" as your results (and it only gets worse with "qualitative data" and qualitative analyses). How anyone swallows this more than naive practice and presentation of results in one gulp (other than Alice) without a little thought and critical reflection is more than just curious, it is actually statistically mystifying. Various authors and experts who make a variety of negative urban legend claims about the Likert response format (which they inappropriately and inaccurately call "scales") are fixated on the single item (or are talking about this mode of data analysis), and these authors and experts confuse the underlying continuum that is the "scale" of the "objective and fused" response format with the underlying continuum of the collection of items that is the "scale" **of the variable being measured**—two very different things with very different properties because of the differing levels (micro versus macro), even though one is or might be talking about the **same scalar properties** (i.e., order, equal units, a true zero point, linearity or lack of these properties). These are Ryle-like category mistakes of the first order with rippling multiplier effects[28]. This category mistake and misunderstanding, along with a lack of knowledge of a number of key empirical facts, leads to perhaps the most widely known erroneous or mythical claim about "Likert scales," which is that "Likert scales are ordinal scales and thus only non-parametric statistical tests may and should be used with them."

F Is Not Made of Glass: The non-parametric statistical analyses only myth about "Likert scales" is particularly disturbing because many (if not all) "item fixated" experts seem to be completely unaware of Gene Glass' famous Monte Carlo study of ANOVA in which Glass showed that **the F-test was incredibly robust to violations of the interval data assumption** (as well as moderate skewing) and could be used to do statistical tests at the scale and subscale (4 to 8 items but preferably closer to 8) level of the data that was collected using a 5 to 7 point Likert response format with **no resulting bias**[29]. Glass also showed that the F-ratio could also actually be used to do **a priori** testing of selected Likert response format items at the item

level if there were a sufficient number of scale points. So Glass showed that the F-test is not made of glass and the the F-test is extremely robust (except to violations of the equality of variances assumption), and that one does not have to lose statistical power and sensitivity by using non-parametric statistical tests in its place when analyzing Likert scale data and even analysis such data **selectively** at the item level.

The underlying conceptual and empirical reasons why Glass found the Monte Carlo results for the F-ratio he found can be explained by the results of the studies done by Carifio of the Likert and other response formats and types of responding continua, where the data were shown empirically to be both linear and interval at the macro (or measurement) scale level[22-24]. So analysis of Likert response format data using the F-test is not only statistically robust at the item level, when the testing is on an a priori (and not shotgun data fishing) basis and the number of scale points is sufficient (preferably 7), but the Likert response format does not have, in fact, the many problems alleged experts claim it has, and these two points are even more true at the macro (measurement) scale level, which is the level the analyses should be occurring at in the first place. Also, if the data is thought not to be interval level data, then the data may be tested to see if "the scale units are equal and/or that the scale itself is linear" is in fact the case[22,30,31]. If the data are not interval or/and linear prior to analysis, they may be scaled to be so if necessary, if the consequences of accepting a false null hypothesis are that great (as opposed to just radically increasing the alpha levels). Scaling the data so it is "perfectly interval data," it should be noted, will improve (in most cases) the Pearson correlation coefficients, which in turn will have very significant multiplier effects of various kinds in all statistical analyses that use the correlation coefficient as one of if not the fundamental unit of input to the analysis (e.g., multiple regressions, factor and discriminant analyses, and the multivariate F-test). So it is really the correlation coefficient that is most effected by "scale" and "data" type, which is the real, core and key problem that is never mentioned or discussed by various experts on Likert scale, Likert response formats, and statistical analyses thereof, with one notable excellent exception[32]. F is not made of glass but correlation coefficients **are** to a great degree, and this particular empirical fact and its many consequences are one of the greatest silences in all of this literature. These empirical facts, therefore, only leaves the problems of misinterpreting the meaning of the "scale" or results obtained with it at either the item or the macro scale level.

The Likert Code: The basic misinterpretation problem with "Likert scales," which has nothing to do with the scale itself, but rather emanates from a misunderstanding of data type and associated psychologies of interpretations, is the interpretation tendencies (and belief) that an anchoring term (or attribution label) such as "agree" is "twice as much" or "one more unit as much" as "somewhat agree" and so on through the possible comparisons that can be made in the response format labeling terms used. This type of interpretation of the data is usually due to inadequate knowledge and logical and interpretive errors, which tend to be negative transfers of prior "scale using" experiences, as an ordinal scale is not an equal unit scale and one cannot make ratios of the responses. One, therefore, doesn't know how much stronger or how many more units "agree" is than "somewhat agree" (not to mention that a "somewhat disagree" response and "somewhat agree" response could actually represent a differing number of units of "agreement"!). But most of these same points are also true at the **macro scale level** even if the data are empirically an interval scale at this level, as in the studies done by Carifio[22-24] and the subsequent studies cited[25-27].

An **interval scale has an arbitrary zero point (not a true zero point) so one cannot make ratios or interpret the data in "ratio ways" when the data is interval either.** Also, if the data are interval at this macro (total instrument/scale) level, then the number of units from somewhat agree is approximately the same as the number from agree to strongly agree, but you do not know how many units exactly because of the type of scale and its metric limitations (but there are statistical procedures that can tell you). In this situation, you still do not know nor can you easily estimate what it would take in terms of treatments, instructional efforts, funding strategies, dosage levels or hours of therapy (i.e., "units of energy") to move a person those many units on the scale; namely, what it would take to get people in general or a person from one anchoring term on the scale to another. It might take changing their opinion 2 categories on 5 of the items or some other parameters requiring a sophisticated analysis to establish. An equal unit scale does not logically or necessarily mean a one to one correspondence between individual items and macro scale units, which is an inappropriate generalization from multiple-choice (dichotomously-scored) achievement scales. So with a few basic facts, some

knowledge, several examples, and a little practice interpreting data and results from a Likert scale and Likert response formats, correcting important measurement misconceptions is not as hard or tough as breaking the DaVinci code.

Quarks: So what exactly is a scale then? One must put the word in context to both define and understand what a scale is and how to operationalize one effectively. At an **empirical minimum** a "scale," at the item or group of items level, must have some underlying continuum and rank ordering points along this underlying continuum; otherwise, it is a nominal classification topology. But this definition and view of a scale is a blindly empirical, mathematical, and meaning-free definition and view of a scale, which is another part of the core and fundamental problems in this area. For all of the reasons discussed above in detail, a scale is much more than this minimum empirical and mathematical definition and view, and a scale cannot be separated from its "semantic" and "meaning" and purportment and property specifications (i.e., instrument blueprint and "tables of specifications"), which is one of the roots of the current problems and inaccurate and untrue claims. One cannot separate the "semantic" (i.e., meaning components and specifications) of a scale from the "grammatical" (i.e., form and mathematical) components of a scale and throw the semantic components away or pretend that they do not exist and are unimportant without great problems, confusions, and difficulties ensuing. But this "separation and discard as inconsequential approach" is exactly what logical positivists, blind empiricist, mathematicians, mathematical psychometricians, mathematical measurement professionals, and untrained (or poorly trained) practitioners do creating many of the problems discussed in this article as well as many more problems.

The "semantic" and the "grammatical" components of a scale really cannot be separated other than "conceptually" as they are a fundamental unit and "whole" like the "quark" and are in fact inseparable and separating them changes and "denatures" them just as frying an egg (permanently) denatures its proteins. Scales at either the macro or micro level, therefore, are not heaps of disconnected organs on a table anymore than a person is such a thing. One cannot logically disembowel a scale and still really have a scale, and when one does disembowel the semantic components of a scale from the grammatical components and focuses solely on the grammatical (i.e., mathematical) components, the "scale" often appears to behave in strange and unpredictable ways. Fortunately, scales are in many ways similar to quarks so that the more you try to pull the pair (semantic and grammatical) apart, the more strongly they bond and fuse and must be considered as one unit. Given these points, we can thus re-ask the question, "What then is a scale?"

A macro scale or measurement scale (as opposed to the underlying response format/coding scale of an item) is a purposely constructed (according to an a priori blue print and plan) inter-related set of items which have defined and targeted logical and empirical properties. An item captures logically predefined "units of information" about the variable and construct being measured. In a completely "open-ended" more "qualitative" model, one, in the end, has a set of predefined units of information that one sieves from the free and unstructured flow, gathering each unit instance up from the "streaming thick and rich information flow (image or text)," as well as the interconnections between them. So the models present here in this article are quite general and not just confined to "Likert scales," nor are many points in the current discussion. The minimum number of items needed to have a measurement scale (or test or subtest) is 6 to 8 items due to various reliabily, validity and generalizability considerations[33]. The scale or subscale, then, is the logical and empirical properties of the items individually and as a whole, or as a collection (subscale), or collection of collections (total scale). Items are suppose to be logically related to the predefined construct the scale is measuring and free from defects that would taint or distort the information the item captured relative to the construct. So an item may be substantively or logically flawed and unacceptable or defective as an item, or "mechanically" flawed, defective and unacceptable as an item.

Arguments, claims and criticisms about Likert Scales tend to be about the Likert item response format (a "mechanical" characteristic), and the response format's alleged mathematical characteristics, and not about the logical characteristics that the items of a Likert Scale must have, because few of the experts, or those making these claims, seem to be aware of this important distinction, or the importance of the logical (and content) properties of instruments and scales. Also, these experts most likely have not read Likert's original book on what a Likert (Attitude/Opinion) Scale is or how one must be constructed[14]. These experts, as well as many others, have detached the Likert response format from the Likert Scale, and this very disembodiment is a key part of the problem and source of the misunderstandings and urban legends that now abound. As stated above, scales can be thought of as quarks and the components of scales cannot be separated or disemboweled and ignored or disregarded, even at the item response format (i.e., micro) level, where the semantic anchoring terms used affect the **scalar properties** (i.e., unit equivalence and so on) of

the captured information and resulting data. Nowhere are these points more true than for a Likert Scale.

The Likert Scale: According to Likert[15], to have a Likert Scale, one had to write a series of verbal statements that expressed a range of positive expressions, views, sentiments, claims or opinions about the "attitude object (underlying construct)" that ranged from mildly positive to strongly positive and then the same relative to a range of negative statements. Logically, someone who is positive about the attitude object should agree with the positive statements and disagree with the negative ones (thus the need for "reverse item scoring") so a logical check and validity is built into the construction protocol (unlike a variety of questionnaires now that use the Likert response format). If the "attitude object" was believed to have some sub-dimensions of importance, this methodology was repeated for each of the sub-dimension or sub-variable or sub-scales, again with 6 to 8 items balanced by positive and negative statements (i.e., items) being the criteria for the subscale. Once such a set of items were developed, they were then given to an independent set of judges to Q-sort into the appropriate categories of positive and negative expressions and degrees of expression (i.e., their essential characteristics).

So the logical properties and criteria of a Likert Scale (or any scale) are one's first and foremost concerns and features. It is these logical features and criteria that produce the variations in total scale score (i.e., allow them to be expressed if there) that would be present with enough items, even if a binary response format was used. To gain efficiencies in measurement, cost and response times (as well as the need to construct fewer perfect items (as authentic Likert items are hard to construct), Likert devised the Likert response (information capturing) format to go with the graded verbal statements of his Likert Scale. As previously stated, the Likert response format **fuses** two of the three logical/conceptual components of an "item" (i.e., the real time response and the scoring/coding/interpretation of the response) into an efficient and cost effective self-coding format which has semantic and grammatical (including mathematical) scalar properties. In this sense, the Likert response format is a "passive/selective" as opposed to a "generative" and "open-ended" response format[34], and the Likert response format is the "multiple choice item of the affective domain."

Three points should be note here. First, there is little if any (conceptual) difference between having an open-ended question that some one responds to in writing or in an interview in some kind of recorded fashion that one then codes with a protocol that has scalar properties built into it and a Likert response format other than the "separation" of the two components and who is doing the coding. Next, there is nothing derogatory about calling the Likert response format the "multiple choice items of the affective domain," if one has a cognitive learning theory view of the multiple choice item and knowledge of all of the things that can be done with this type of item response format using this model and theory of the multiple choice item. Lastly, the central point here is that the Likert response format worked for Likert and his Likert Scale because it was connected to each statement on the purposefully constructed Likert Scale and the statement's logical requirements and characteristics.

It was other researchers who disconnected the Likert response format from the purposefully constructed Likert Scale with its key logical and semantic requirements (which are very difficult in fact to meet and one reason why the "detachment" occurred and few Likert Scales are now constructed). But in reality the Likert response format and scale cannot be "detached" because the characteristics (logical and semantic) of the questions, items or statements in the questionnaire, test, or scale are key and drive and determine everything. This basic fact is what "blind shotgun empiricist" researchers and commentators who write articles in this area and on this topic simply do not understand or ignore, as logic and meaning are irrelevant in their view, particularly as compared to the "numerical/empirical properties." Logic and meaning are the only and very thing that make numerical and empirical properties interpretable, which is the key and central points all of these experts miss and which we want to emphasize clearly here. All of these points become readily apparent when one realizes that factor analyses (a structural and content-based construct validity model) are now used routinely to empirically assess the quality and characteristics of scales and questionnaires that use the Likert responding format[35,36].

It should also be noted the actual questions that comprise a Likert Scale can themselves be scaled to add further refinements and weighted scoring to the aggregation of items into subscale and total scale scores, which also tends to improve the linear and interval scale properties of the resulting composites. One simply cannot detach a response format from the content and logic associated with it or ignore the content and logic of an instrument (collection of items) or build a psychometric theory or develop answers to psychometric questions to the single item in isolation from all else; that is blind, shotgun, and disconnected and detached empiricism which is the nub of the problem and the problems strewn throughout this literature.

Table 1: The Top Ten Myths and Urban Legends about "Likert scales" and the Counter Argument and "Antidote" for Each Myth and Urban Legend.

Myth 1—There is no need to distinguish between a scale and response format; they are basically the same "thing" and what is true about one is true about the other. **ANTIDOTES: SCALES VERSUS RESPONSE FORMATS, TSS (TECHY SLOVENLY SLANG) IPES (INTELLECTUAL PIGEON-ENGLISH SPEAK), ATOMS, MOLECULES AND SCALES.**
Myth 2—Scale items are independent and autonomous with no underlying conceptual, logical or empirical structure that brings them together and synthesizes them. **ANTIDOTES: ATOMS, MOLECULES AND SCALES, NEO-POSITIVISM AND UNBLIND EMPIRICISM, AND MODERN SCHEMA THEORY.**
Myth 3—Likert scales imply Likert response formats and vice versa as they are isomorphic. **ANTIDOTES: LOGICALLY INCORRECT, NEITHER ARE NECESSARY OR SUFFICIENT FOR THE OTHER, MACRO AND MICRO LEVELS OF MEASUREMENT, ATOMS, MOCELULES AND SCALES.**
Myth 4—Likert scales cannot be differentiated into macro and micro conceptual structures. **ANTIDOTES: LOGICALLY INCORRECT, MACRO AND MICRO LEVELS OF MEASURMENT, THE QUESTIONS ARE THE NUB.**
Myth 5—Likert scale items should be analyzed separately. **ANTIDOTES: ONE SWALLOW, ONE ITEM A SCALE DOTH NOT MAKE, FAMILY-WISE ERROR RATES, ATOMS, MOLECULES AND SCALES.**
Myth 6—Because Likert scales are ordinal-level scales, only non-parametric statistical tests should be used with them. **ANTIDOTES: F IS NOT MADE OF GLASS, STUDIES SHOW LIKERT RESPONSE FORMATS TO BE INTERVAL SCALES, LIKERT RESPONSE FORMATS MAY BE RATIO SCALES LOGICALLY WITH THE CORRECT ANCHORING TERMS.**
Myth 7—Likert scales are empirical and mathematical tools with no underlying and deep meaning and structure. **ANTIDOTES: QUARKS, THE QUESTIONS ARE THE NUB AND THE LIKERT CODE.**
Myth 8—Likert response formats can without impunity be detached from the Liker Scale and its underlying conceptual and logical structure. **ANTIDOTES: QUARKS, FUSION AND THE LIKERT CODE.**
Myth 9—The Likert response format is not a system or process for capturing and coding information the stimulus questions elicit about the underlying construct being measured. **ANTIDOTES: QUARKS, THE QUESTIONS ARE THE NUB, FUSION AND THE LIKERT CODE.**
Myth 10—Little care, knowledge, insight and understanding is needed to construct or use a Likert scale. **ANTIDOTES: ALL ANTEDOTES PLUS MEASUREMENT, TRAINING AND RESEARCH LITE.**

The Questions Are the Nub: Every test, questionnaire, instrument, interview protocol, scale, and so on has a stimulus component and a response component and a context component. Likert, like Cronbach[37] and a long list of other test and psychometric experts focused first and strongly on the stimulus component (namely, the construct to be measured and the set of logically interrelated "items" that would elicit desired units of information about the construct in question)[38,39]. This focus and fact is the essentially missing ingredient in the blind empirical or Thorndikian approach to measurement, and so much of what is done and said currently. Blind empiricism only produces blind data and blind results. Scales are about the macro collection of these information eliciting items, which is the place where the first and foremost focus should be put.

The response component of this model concerns (1) capturing the information elicited and (2) transforming what is captured into the meaningful "units of analysis" (scales and subscales) that are the fundamental building blocks of analyses and hypothesis/theory testing or "answering the research questions," which in research slang is called "scoring your instrument" or your "scoring protocol." The Likert response format is a technology for capturing information the stimulus questions elicit. As previously stated, the Likert response format has two components (logical/semantic and mechanical) and many of its alleged defects and problems are vastly over-stated and based on flawed and mistaken logical arguments or assertions only. Further, these arguments and claims tend to embody category mistakes and conceptual flaws and are contradicted by well-established empirical facts as well as careful (and less sloppy) logical arguments, analyses and conceptual and theoretical views and models. A Measurement Scale is something very different from and much more than just an

underlying continuum of a response format for one or more individual items. This category mistake and error is one of the first-order, with incredible ramifications and consequences, as one can easily see from reading the literature in this area and on this subject. The emergent property of a measurement scale (collection of items) is not the same thing as the property or underlying scale of each of the items that constitute the emergent scale in the same way that the properties of molecules are very different from their constituent atoms. The alleged problems of Likert response formats are all testable and correctable and the F-ratio and many other parametric statistics and tests are incredibly robust to violations of the interval scale assumption. Further, this latter point was empirically settled over 40 years ago, as were questions and issues about the underlying continuum of the Likert response format.

Measurement, Training, and Research Lite: The problem, therefore, is really a problem of extremely poor and careless scale (test, questionnaire, interview, protocol and so on) construction today, and even poorer and more careless data analysis than it is a problem of any inherent, conceptual, untestable or uncorrectable problem with the Likert response format itself. The list of claims made by Jamieson[1] (and all of the various authorities she cites, as well as others) about Likert response formats (with nary a word about actual and true Likert scales) are not only misconceptions, misunderstandings, category mistakes, and logically incorrect and untrue, but these claims and assertions are also empirically inaccurate and untrue and mainly based on armchair analyses and speculations and an unacceptable level of carelessness relative to established measurement and psychometric theory and research, and what Likert himself actually had to say about his own inventions and what is known about them empirically. Further, these later points are even truer relative to the interpretation of results by many researchers who know little about this research tradition in depth, or what the measurement scale values that result from use of the Likert response format mean at the macro level, never mind for the individual (micro level) items. And it is these same researchers, specialists and practitioners who persist, despite the voluminous number of logical and empirical arguments to the contrary, to analyze their data at the item level and to present their findings as long unorganized laundry lists of item-by-item results. Such practices could be called the "cult of the individual item and subatomic micro analyses at the twig never mind tree level." However, this item-by-item (fuzzy jumble) analysis practice and strategy simply has to stop, and particularly so with two and three category response formats, except if the analysis is an item analysis or formative and exploratory data analyses (or the very best that can be done in a very limited situation), which, as previously stated, is not where the analyses stop or are presented. If one is using a 5 to 7 point Likert **response format**, and particularly so for items that resemble a Likert-like **scale** and factorially hold together as a scale or subscale reasonably well, then it is perfectly acceptable and correct to analyze the results at the (measurement) **scale** level using parametric analyses techniques such as the F-Ratio or the Pearson correlation coefficients or its extensions (i.e., multiple regression and so on), and the results of these analyses should and will be interpretable as well. Claims, assertions, and arguments to the contrary are simply conceptually, logically, theoretically and empirically inaccurate and untrue and are current measurement and research myths and urban legends. It is more than time to dispel these particular research myths and urban legends as well as the various damage and problems they cause, and put them to bed and out of their misery once and for all.

CONCLUSION

In an effort to summarize this article and its many points, a list of the current myths and urban legends identified here and the counter arguments or "antidotes" to these myths and legends is given in Table 1. As can be seen from Table 1, the section headings of this article and certain key concepts (Table 1 does not include just the section headings) are given as the antidotes and counter arguments for each myth and one may reread the sections given and assemble all of the details, arguments, and empirical evidence needed to refute the myth in question. In conclusion, then, particularly after reading and evaluating the summary given in Table 1, one should be able to see that "the questions are the nub" and "one item a scale doth not make;" that there are "macro and micro levels of measurement", and critical differences between measurement "scales and response formats;" that "the F-test is not made of glass"[29], and there is a "Likert Code," in reference to what "Dr. Likert actually said and did"[15], that the "semantic and grammatical/mathematical" components of items and (measurement) scales are like "quarks" and really cannot be pulled apart and discarded or ignored (and particularly so the semantic components); that Likert response formats can empirically produce interval[22,23] and even (for sake of argument and to make a point) ratio data logically and empirically[27]; that measurement, training and research Lite are not good enough or acceptable anymore, and that the practice of analyzing "Likert scale" questions item-by-item and presenting the results the same way, and as an unorganized laundry list and fuzzy jumble (whether done quantitatively or qualitatively) must simply stop as a research and reporting practice; and that the various persistent myths and urban legends about Likert scales and Likert response formats must be eliminated once and for all through better education and training.

REFERENCES

1. Jamieson, S., 2004. Likert scales: how to (ab)use them. Medical Education, 38: 1212-1218.
2. Knapp, T.R., 1990. Treating ordinal scales as interval scales: an attempt to resolve the controversy. Nursing Research, 39: 121–123.
3. Kuzon, W.M Jr., M.G. Urbanchek and S. McCabe, 1996. The seven deadly sins of statistical analysis. Annals of Plastic Surgery, 37: 265–272.
4. Pett, M.A., 1997. Non-parametric Statistics for Health Care Research. London, SAGE Publications.
5. Cohen, L., L. Manion, and K. Morrison, 2000. Research Methods in Education. 5th ed. London, Routledge Falmer.
6. Blaikie, N., 2003. Analyzing Quantitative Data. London, Sage Publications
7. Thurstone, L.L., 1954. The measurement of values. Psychological Review, 61: 47-58.
8. Siegal, S., 1956. Non-parametric Statistics. New York, Wiley.
9. Coombs, C.H., 1960. A theory of data. Psychological Review, 67: 143-159.
10. Torgeson, W.S., 1961. Scaling and test theory. Annual Review of Psychology, 12: 51-70.
11. Mehrens, W. and R. Ebel, 1967. Principles of Educational and Psychological Measurement. Chicago, Rand-McNally.
12. McNemar, Q., 1969. Psychological Statistics. New York, Wiley.
13. Summers, G.F., 1970. Attitude Measurement. Chicago, Rand McNally.
14. Likert, R., 1932. A Technique for the Measurement of Attitudes" Archives of Psychology 140, 55.
15. Likert, R. and S. Hayes, 1957. Some Applications of Behavioural Research. Paris, Unesco.
16. Campbell, D.T. and J.C. Stanley, 1963. Experimental and Quasi-Experimental Designs for Research. New York, Rand-McNally.
17. Uebersax, J.S., 2006. Likert scales: dispelling the confusion. Statistical Methods for Rater Agreement. Available at: http://ourworld.compuserve.com/homepages/jsuebersax/likert2.htm.
18. Kerlinger, F. and H. Lee, 2002. Foundations of Behavioral Research (4th ed.). New York, Harcourt.
19. De Vellis, R.F., 1991. Scale Development: Theory and Applications. London, SAGE.
20. Fife-Schaw, C., 1995. Levels of measurement. In G. M. Breakwell, S. Hammond and C. Fife-Schaw (Eds.) Research Methods in Psychology. London, SAGE.
21. Coolican, H., 1999. Research Methods and Statistics in Psychology, 2nd ed. London, Hodder & Stoughton.
22. Carifio, J., 1976. Assigning students to career education programs by preference: scaling preference data for program assignments. Career Education Quarterly, 1, 1, Spring, 7-26.
23. Carifio, J., 1978. Measuring vocational preferences: ranking versus categorical rating procedures. Career Education Quarterly, 3, 1, Winter, 34-66.
24. Carifio, J., 1990. Preparing teachers attitudes towards tests. Paper presented at the annual conference of the Eastern Educational Research Association, Tampa, FLA, ERIC TM014483.
25. Baggaley, A. and A. Hull, 1983. The effect of nonlinear transformations on a Likert scale. Evaluation & the Health Professions, 6: 483-491.
26. Mauret, J. and H.R. Pierce, 1998. A comparison of Likert scale and traditional measures of self-efficacy. Journal of Applied Psychology, 83: 324-329.
27. Vickers, A., 1999. Comparison of an ordinal and a continuous outcome measure of muscle soreness. International Journal of Technology Assessment in Health Care, 15: 709-716.
28. Ryle, G., 1949. The Philosophy of Mind. London, Oxford Press.
29. Glass, G.V., P.D. Peckham and J.R. Sanders, 1972. Consequences of failure to meet assumptions underlying the analyses of variance and covariance. Review of Educational Research, 42: 237-288.
30. Edwards, A.L., 1956. Techniques of attitude scale construction. New York, Appleton-Century-Croft.
31. Bock, R. and L.V. Jones, 1968. The measurement and prediction of judgment and choice. San Franciso, Holden-Day.
32. Russell, C.J. and P. Bobko, 1992. Moderated regression analysis and Likert scales: Too coarse for comfort. Journal of Applied Psychology, June, 77(3): 336-342.
33. Cronbach, L,J., G.C. Gleser, H. Nanda and N.S. Rajaratnam, 1972. The Dependability of Behavioral Measurements. New York, Wiley.
34. Nasser, R. and J. Carifio, 1994. Key contextual features of algebra word problems: a theoretical model and review of the literature. Paper presented at the annual conference of the Eastern Educational Research Association. February, ERIC SE 053490
35. Sisson, D.A. and H.R. Stocker, 1989. Analyzing and interpreting Likert-type survey data. The Delta Pi Epsilon Journal, 31(2): 81-85.
36. Clason, D. and T. Dormody, 2002. Analyzing data measured by individual Likert-type items. Journal of Agricultural Education, 35(4): 31-36.
37. Cronbach, L., 1990. Essentials of Psychological Testing (5th ed.). New York, Harper-Row.
38. Nunnally, C, and I. Berstien, 1994. Psychometric Methods. New York, McGraw Hill.
39. Guilford, J., 1954. Psychometric Methods (2d ed.). New York, McGraw Hill.