Title

Seeking a Better Balance Between Efficiency and Interpretability: Comparing the Likert Response Format With the Guttman Response Format

Authors

Wilson, Mark
Bathia, Shruti
Morell, Linda
et al.

Publication Date

2022-01-13

DOI

10.1037/met0000462

Peer reviewed

HHS Public Access

Author manuscript

Psychol Methods. Author manuscript; available in PMC 2023 July 13.

Seeking a better balance between efficiency and interpretability: Comparing the Likert response format with the Guttman response format

Mark Wilson[1], **Shruti Bathia**[1], **Linda Morell**[1], **Perman Gochyyev**[1], **Bon W. Koo**[2], **Rebecca Smith**[3]

[1]Graduate School of Education, University of California, Berkeley

[2]Lawrence Hall of Science, University of California, Berkeley

[3]University of California, San Francisco

Abstract

The Likert item response format for items is almost ubiquitous in the social sciences, and has particular virtues regarding the relative simplicity of item-generation, and the efficiency for coding responses. However, in this paper, we critique this very common item format, focusing on its affordance for interpretation in terms of internal structure validity evidence. We suggest an alternative, the Guttman response format, which we see as providing a better approach for gathering and interpreting internal structure validity evidence. Using a specific survey-based example, we illustrate how items in this alternative format can be developed, exemplify how such items operate, and explore some comparisons between the results from using the two formats. In conclusion, we recommend usage of the Guttman response format for improving the interpretability of the resulting outcomes. Finally, we also note how this approach may be used in tandem with items that use the Likert response format to help balance efficiency with interpretability.

Keywords

Likert scales; Likert response format; Guttman response format; internal structure validity evidence

In a tribute to its wide-spread usage, Wikipedia describes the Likert Scale as “the most widely used approach to scaling responses in survey research, such that the term … is often used interchangeably with rating scale” (Wikipedia entry for “Likert Scale”). We see two bases for this popularity, (a) the ease of item development and the efficiency of the item responses format, and (b) how the item scores are then summed to calculate a total score

Correspondence concerning this article should be addressed to Mark Wilson, Graduate School of Education, University of California, Berkeley. MarkW@berkeley.edu.

We have no known conflicts of interest to disclose.

As noted in the text, parts of the data and results relating to the Likert-style items and the Guttman-style items have been reported earlier, in Bathia et al. (2020) and Morell et al. (2021), respectively. However, the comparison of the two, which is the focus of this paper, has not been reported before.

(Likert, 1932/3). In this paper, we focus on the former aspect, the so-called *Likert response format* (Carifio & Perla, 2007), which is also described in Wikipedia:

> "A Likert item is simply a statement that the respondent is asked to evaluate by giving it a quantitative [or ordinal] value on any kind of subjective or objective dimension, with level of agreement/disagreement being the dimension most commonly used." (Wikipedia entry for "Likert Scale")

While concurring with the general appraisal of the efficiency of coding responses, and with the relative simplicity of item-generation under this scheme, in this paper, we consider also criticisms of the Likert response format for items, and, in particular we focus on concerns about relating the response categories of the items to the structure of the construct—a validity issue, usually termed "evidence of internal structure" (AERA/APA/NCME, 2014)—and hence, also with associated limitations in interpretability of the resulting scale. In light of those criticisms, we suggest an alternative format, the *Guttman response format.* We illustrate how items in this format can be developed, explain how such items operate using a specific survey-based example, and explore some comparisons between the results from using the two formats. In conclusion, we recommend usage of the Guttman response format for improving the interpretability of the resulting outcomes. However, we also note how this approach may be used in tandem with items that use the Likert response format to help balance efficiency with interpretability.

The Likert response format for items

Although Likert (1932/33) is, of course, the most authentic source for the definition of Likert-style items, at the time, Likert himself was not actually aware that he was establishing a "brand-name" that would become so popular in the ensuing decades, and so was not strongly motivated to provide a clear definition of the term. A recent textbook by DeVellis (2017) gives a clear and succinct description as follows:

> "the item is presented as a declarative sentence, followed by response options that indicate varying degrees of agreement with or endorsement of the statement." (p. 93).

However, the concept has very wide usage, and hence is mentioned, described and defined in multiple papers and textbooks, as well as having a large presence on the internet (see, for example, Robinson (2018) or Uebersax (2006)). Thus, one can recognize the example items in Figure 1 as fairly typical examples of the genre. Sometimes the format is varied by the item developers—for example, there might be no word labels, just numbers, or there might be a line, which is labelled at equally-spaced intervals with numbers. A fairly common alternative is the "semantic differential" (Osgood & Tannenbaum, 1955), where only the ends are labelled.

It is important to note that this paper is focused on the response format of items, and especially, consequences of that choice for interpretation and the gathering validity evidence. However, it is common in the literature to see a confusion between, for example, references to a Likert *scale*, and the Likert *response format*. While acknowledging that the word "scale" has multiple meanings (especially within the measurement domain), we emphasize that the

main focus of this article is the response format rather than the resulting "scale" based on the set of items. For an interesting (though somewhat idiosyncratic and testy) discussion of this distinction see Carifio and Perla (2007).

Some criticisms of Likert-style items.

There have been numerous criticisms of Likert response format items over the almost 100 years since Likert wrote his foundational paper (as one might expect for anything that is so common). Among them are criticisms that there can be a tendency for people to answer on only one response side or the other, a tendency for people to not choose extremes, or that there is a confusion between an "equally-balanced" response and a "don't know/ does not apply" response (DeVellis, 2017). However, probably the most common criticism by psychometricians of Likert-style items is that the use of integers for coding the responses implies that the responses are equal-interval, seemingly conferring (at least) interval-level measurement status on the resulting data, and hence that statistical procedures requiring such (for example, linear regression, factor analysis, etc.) can be confidently carried out. In fact, it is very common that such data-analytic procedures are carried out, despite strong criticisms of such practices over many years (Carafio & Perla, 2007; Jamieson, 2005; Kuzon et al., 1996; Ubersax, 2006). However, there is also a long literature on the robustness of such procedures against such violations, dating back even to Likert (1932/33) himself, but with other commentators agreeing over the years (*cf.,* Glass et al., 1972; Labovitz, 1967; Traylor, 1983).

However, this is not the prime criticism to be made here, but we will return to it at a certain point later in the paper, to try and throw some light on this controversy. The reason that this debate is not considered in this paper is that the approach that will be taken here is the modern psychometric perspective that the numerical values for the Likert responses themselves do not constitute the measurement scale, but rather that these are observations of discrete-valued random variables (i.e., one random variable for each item) that is to be modelled as being probabilistically dependent on an underlying latent variable. It is this underlying latent variable which we take to constitute property which will be labelled with the measurement scale (for the set of items), and that this latent variable has its own scale-structure, which we will take to be an interval scale. This strategy obviates the not-equal-interval criticism of the raw Likert responses (see similar comments, for example by Embretson (1996) and Wright and Masters (1982).

Establishing internal structure validity evidence

In this paper we will focus on those aspects of a scale that relate most directly to its meaningfulness, internal structure validity evidence, and the interpretation of the results from use of the instrument. The specific logic for the sort of internal structure validity that we are addressing here is that the structure that is posited as belonging to the property being measured should be reflected in empirical evidence available once data has been collected, and the results evaluated (AERA/APA/NCME, 2014). To make this case, it is first required that there be an intended (and evaluable) structure of the property being measured, and this is something that should be included in the content validity evidence, that is, there needs to

be a structural description of the property under measurement, often called the "construct" in educational and psychological measurement. In the case of a unidimensional property, this corresponds to a description of the property that can be realized either through a structural account of the respondents or the items (or both).

One prominent version of such an account is conceptualized as the *construct map* (Wilson, 2005). In this approach, we need a particular sort of description of the construct—first, assume that the construct we wish to measure has a simple unidimensional form: it extends from one extreme to another, from, say, high to low, or strong to weak, with qualitatively distinct locations in between. We are primarily interested in finding where a respondent stands on this range from one extreme to the other (i.e. we wish to measure the respondent). In particular, when we are relatively more sophisticated about our understanding of the construct, it may be possible to distinguish a sequence of qualitative levels between those extremes—these will be very important and useful both for validity evidence, and for the interpretation of the measurements. At this point, it is still an idea, latent rather than manifest, and although qualitative distinctions within the range should be definable, we assume that the respondents can be at any point in between—that is, the underlying construct is continuous.

An example construct map

We next will show an illustration of a construct map from an example of a Likert Scale development which will be used throughout the paper. First, we give some brief background on this instrument to make the example clearer and more concrete. The instrument was developed to measure researcher identity by the San Francisco Health Investigators (SFHI) Project (Koo et al, 2021) and is referred to as the *Researcher Identity Scale* (RIS). The developers considered RIS to be one unified idea made up of four strands: fit & aspiration, community, self, and agency. For more information on the project under which this development work took place, see Koo et al. (2021) and for more information about the RIS latent variable, its components, and its construct map, see Bathia et al. (2020). The construct map for this construct is shown in Figure 2. The hypothesis of the RIS construct map starts at the lowest level (Level 0), below the levels shown in the construct map in Figure 2, where the student is not aware of what research entails and has no consideration for their possible role(s) in research. At Level 1, the student is a newcomer to the concept of research. At Level 2, the student explores the different aspects of research. At Level 3, the student begins to feel comfortable with their identity as a researcher. And finally, at Level 4, the student identifies themself as a researcher and integrates this into their larger self-identity.

With a construct such as this in hand, the specific goal of an internal structure validity investigation will be to find evidence about whether this structure is reflected in the data collected in a validity study of the instrument designed to measure the latent variable (Wilson, 2005)—in this case the RIS latent variable. Following the typical development steps for attitude scale construction, the RIS developers created items following the Likert response format approach. Examples of the items they developed have been shown in Figure 1. The SFHI Project developed a 45-item Likert response format instrument with six

response categories for each item as shown in Figure 1 (i.e., Strongly Disagree, Disagree, Slightly Disagree, Slightly Agree, Agree, and Strongly Agree).

The Guttman response format for items

An alternative to the Likert response format is to develop options that are in themselves meaningful statements that give the respondent some context in which to make the desired distinctions. The general aim of such a tactic is to try and make the relationship between each item and the overall scale interpretable: the specific aim for this study is to try and make the relationship between each item response option and the construct map levels interpretable. If we can be successful at the latter, then the former will follow as part of the results from data analysis. On the basis of items such as these, Guttman developed his very unique and intuitive approach to measurement (1944), which he called his "scalogram" approach (also known as "Guttman scaling"):

> If a person endorses a more extreme statement, he should endorse all less extreme statements if the statements are to be considered a [Guttman] scale… We shall call a set of items of common content a scale if a person with a higher rank than another person is just as high or higher on every item than the other person. (Guttman, 1950, p.62)

Four items developed by Guttman himself using this approach are shown in Figure 3—these items were used in a study of American soldiers returning from the Second World War (Guttman, 1944). Guttman's own emphasis was on the process of developing scales rather than on items, so he did not name them as a specific type of item. But his accomplishment has been posthumously recognized, and hence this item format design was named as "Guttman-style" items in his honor in 2005 (Wilson, 2005)[1].

The essential feature that characterizes a Guttman-style item is that it provides response options that represent progressively more difficult or higher levels of the construct—that is, not just a generic "stronger," "easier" etc., but options that are themselves meaningfully related to the underlying construct. According to this approach, a deterministic interpretation (such as was made by Guttman himself) would be that the respondent would endorse all the options up to a certain point and then not endorse options beyond that point. That critical point would vary by respondent (i.e., according to their "amount" of the latent variable) and the transduction in the case of the Guttman response format, therefore, would be captured as respondent making their best choice among the ordered set of options. For example, consider the first item in Figure 3. The stem is: "If you were offered a good job, what would you do?" The first option is "I would take the job", and the second option is "I would turn it down if the government would help me to go to school." Note how (a) the options are not generic terms, but are specifically related to the stem, and (b) the Guttman interpretation would be that as a soldier varied from being very interested in getting a job to being very interested in going back to school, there would be a certain point on that (latent) variable where getting government help for school costs was critically important.

[1]Note that this type of item was not invented by Guttman, it had been in used in many previous instruments. In fact, Likert (1932/3) included items like this in his original account on Likert Scales, but they were not taken up by posterity.

A similar interpretation can be made for the comparison of the second to the third option: "I would turn it down and go back to school regardless." The second and third items have stems that are very similar to the first, so that the options for these are also the same or very similar—this would not necessarily be the case for all items in a survey—as the stems varied the contextualization, the options would need to differ to be relevant to those differing contexts. Now, the construct map is a representation of a latent variable that is consistent with this idea of the Guttman response style. Each consecutive level of the construct map can be thought of as representing possible responses that a respondent might make to a certain item stem—in fact this way of thinking of the responses (i.e., the Guttman response format) seems almost tailor-made to fit with the idea of a construct map.

Note that there are two possible orderings that can be considered for polytomous items: the ordering of the items and the ordering (within each item) of the options. This is also illustrated by the items in Figure 3—here the stems (and hence each item as a whole) are ordered, indicated by the type of job—a "good job," "some kind of job" and "no job at all." This ordering in terms of the items is not the prime focus of this paper but is indeed also a possibility that *might* be exploited to reflect the order of a construct map. In the case of dichotomous items, there is no problem with that, but in the case of polytomous items, the conflation (interaction) of the two orders will likely complicate matters with respect to the construct map perspective, and so we have not included this as part of our focus here.

An example of such a series of options in the Guttman response format is shown in Figure 4: this is, in fact, a RIS Guttman response format item. The options in this item are designed to match the levels of the RIS construct map (Figure 2), starting with option (a) for level 0, and progressing through to option (e) for level 4.

Creating Guttman response format items for the RIS scale

Initially, the Likert response format items (such as in Figure 1) were developed to match with levels of the RIS construct map. Then the full set of Likert items were reduced to 21 on the basis of standard quality control indices (appropriate levels of difficulty, etc.). These were then grouped together based on similarity of their content and their match to the construct map levels, to form Guttman-style sets of ordered response options. Not every group mapped across the entire set of construct levels, so some new options were created to fill the gaps. The Guttman-style response options were placed in order based on (a) the theoretical levels of the construct map that they were intended to map to, and (b) empirical evidence of how students responded to the items in earlier rounds of testing. To see an example of this, compare the three Likert-style items shown in Figure 1 with options (c), (d) and (e) for the Guttman response format item in Figure 4. This is, in fact, the matched set of three Likert response format items with one Guttman response format item. As there were not matching items for the two lower levels in the Likert set, two more options were developed for the Guttman item—options (a) and (b) in Figure 4.

Altogether 12 Guttman response format items were developed and validated for the RIS scale (Morell et al., 2021). Eleven of the 12 have at least one Likert-style option that the Guttman options were designed to match to, with 21 matching levels in all, out of a total

possible 60 across all 12 Guttman response format items. Details of the matching of the 21 Likert-style items with the 12 Guttman-style items is given in Appendix A.

A measurement model for both Likert responses and Guttman responses

In particular, we will utilize a Rasch model (1960/80) approach. This is a probabilistic model, and hence, is different from the way in which Guttman discussed his conceptualization of the underlying process (Guttman, 1944). He thought of the critical point mentioned above as determining exactly the response that would occur for each respondent—below that point, the respondent would give one response, at that critical point the response would change to the next option, and so on. This has been a requirement that has not been well-met by the vicissitudes of data in education. For example, extensive investigations by Kofsky (1966) applying Guttman scaling to child development data, led her to observe the following.

> … the scalogram model may not be the most accurate picture of development, since it is based on the assumption that an individual can be placed on a continuum at a point that discriminates the *exact* [emphasis added] skills he has mastered from those he has never been able to perform. … A better way of describing individual growth sequences might employ probability statements about the likelihood of mastering one task once another has been or is in the process of being mastered. (Kofsky, 1966, pp. 202-203)

We agree with this conclusion, and thus, we have chosen a probabilistic formulation. Specifically, we use a polytomous version of the Rasch model, the partial credit item response model (PCM; Wright and Masters, 1982), which is given by:

$$\eta_{pik} = \theta_p - (\delta_{ik}) \tag{1}$$

where $\eta_{pik} = \log \frac{P_{nik}}{P_{nik-1}}$ is the log odds of a person responding in category k versus category k-1 on item i, and where

P_{nik} is the probability of a student responding in category k on item i,

δ_{ik} is difficulty for item i step k, and

θ_p is the location of respondent p.

For, example in the case of a four-category item, with three-steps between categories:

when k=1, Equation 1 defines the log-odds of scoring a 1 rather than 0 for item i;

when k=2, it defines the log-odds of scoring a 2 rather than 1 for item i; and

when k=3, it defines the log-odds of scoring a 3 rather than 2 for item i.

Note that, for the partial credit model, the distances between the consecutive steps do not need to be the same (Wright & Masters, 1982). We will not discuss estimation and technical matters in detail—the interested reader can consult Wilson (2005) or Wright and Masters (1982).

The SFHI Project carried out a study, involving 863 high school students in a western US region, of the reliability and the validity evidence for the RIS instrument and reported results in terms of the first four strands of validity evidence plus fairness, according to the criteria outlined in the "Standards" (AERA/APA/NCME, 2014)—they did not gather any evidence regarding the consequences strand, as there had as yet been no consequences of using the instrument. The data were analyzed using the *ConQuest* software (Adams et al., 2020).

Internal validity evidence for the RIS scale was examined using a graphical representation of the results for the PCM called a Wright Map. An example Wright Map is shown in Figure 5. (This is actually based on data from the Likert RIS data, and we will return to discuss specific results and interpretations for that data in the next section—for now, we use the Figure to describe the features of the graph and its interpretation). A Wright Map is a visual depiction of how the sample of students relate to the items in terms of the PCM parameters for each, respectively. On the far left in Figure 5 one can see the units of the logit scale, which is also represented as the vertical dashed line somewhat off-center towards the right of the Figure. The left side of the Wright map shows the distribution of students in the sample, ranging from those with low estimates at the bottom to those with high estimates at the top in the form of a histogram rotated through 90 degrees so it is "on its side" and where each "x" represents 2.0 students (approximately). The right side of the Wright Map shows the item estimates in terms of their "Thurstonian thresholds" (Adams et al., 2020), which are defined as follows.

a. For an item with the maximum score k (with scores running 0, 1, …, k), there are k Thurstonian thresholds.

b. The kth threshold can be interpreted as the point at which the probability of a score k and above is equal to the probability of scores below k (and hence, both are equal to 0.50 at that point).

All the items in this example[2] have 6 categories, hence each has a maximum of 5 thresholds. In the way that these thresholds are constructed, it is always more difficult for students to reach a higher threshold than a lower threshold (i.e., threshold k is always more difficult to reach than threshold k-1). In Figures 5 the '1s' correspond to the threshold difficulty of moving from "Strongly Disagree" to "Disagree," the '2s' correspond to the threshold difficulty of moving from "Disagree" to "Slightly Disagree" and so on. The different item thresholds are shown in columns on the right-hand side, with labels "Cn" at the bottom representing the item order within the set of items. Further, the item threshold estimates are in the same logit units, thus making it possible to compare item estimates with person estimates, and then, using Equation 1, calculate the probability of a given response on an item using the student location estimate and the item parameter estimates. In this paragraph, we have illustrated the use of the PCM results and the Wright Map for the Likert response format items, but the same description will hold for the Guttman response items, with the difference that here are 5 categories for the Guttman format

[2]I.e., the RIS Likert instrument.

Examining internal validity for the RIS scale

The Likert response format items.

One crucial criterion for internal structure validity evidence is the match between the expectations built into the construct and patterns of results in the instrument outcomes. Thus, an immediate issue in matching the results for the instrument with the levels of the construct map is that, while the construct map levels are described in terms of substantive information about the construct itself, the response categories to the Likert response format items are expressed in terms of amount of agreement (from Strongly Disagree to Strongly Agree). While it is reasonable to assume that these will be related (based on the content of the stems for each item), the actual level of agreement (e.g., Strongly Agree versus Agree for any given item, etc.) that would match to each construct map level in Figure 2 is not clear from a comparison of the contents of the construct map levels (as in Figure 2) with the Likert response options (Strongly Disagree to Strongly Agree). However, an empirical match would be a reasonable step forward, even if the content-matching does not work.

Thus, it makes sense to look for consistency among the Thurstonian thresholds for the Likert response format items, as this could then be a basis for an alignment, perhaps based on a judgment process. As noted above, Figure 5 shows the Wright Map based on the Likert response format items for the RIS scale—the thresholds shown on the right-hand side are just those for the items in the Community section of the instrument. (We have shown just the five Community items in Figure 5—the pattern of results that we are reporting below is the same across all of the topics, and the Figure would have been more cluttered if we showed all 45 items).

What we need to do to is empirically corroborate the construct map levels with the empirical evidence in the shape of the item threshold estimates. To do that, one needs to be able to discern a pattern where each of the sets of thresholds (i.e., the 1st thresholds, the 2nd thresholds, etc.) reside within distinct sections of the RIS logit scale with little or no overlap (Schwartz, et al, 2017). It is readily apparent that, for these results shown in Figure 5, that is not possible, at last without having to accept a lot of overlap between the bands[3]. This is obvious, for instance, by examining the range of the 1st thresholds, which runs (approximately) from −3.4 to −1.0 logits—this range includes within it three 2nd thresholds and one 3rd threshold. This pattern of overlap is repeated for each of the other sets of thresholds except for the 5th. This overlapping pattern of threshold dispersion means that there is no clear one-to-one relationship between the levels of the construct map and the locations of the sets of responses to the items. Now, it may be possible for a creative interpreter to come up with a “theory” about how the item-responses relate differentially to the construct map levels, but this will lead to more exceptions than rules[4]. Thus, it is not possible to establish a case for these results as contributing to internal structure validity evidence with a pattern of results like this. Of course, a wily validity evidence gatherer

[3]To see an example of a Wright map where the bands are much clearer, look ahead at Figure 6 (details in the next section).

[4]For example, suppose that the top of the first band were chosen to be located at −1.5 (just below the 3rd threshold for C1—in order to avoid what looks like a difficult match), then one could decide that what matched to Construct map level 0 (i.e., the lowest) would be “Strongly Disagree,” “Disagree” and (part of) “Agree” for item C1, “Strongly Disagree” and (part of) “Disagree” for item C2 (although a part of “Disagree” would also relate to Construct map level 1), etc.

still has options—this strand of validity evidence might simply be ignored, or some other perspective might be focused on, or, even worse, the idea of being able to describe the qualitative nature of the RIS latent variable might be dropped. However, regardless of these other strategies, it is certainly clear that, for the Likert response format item set, the internal construct validity evidence is not available.

Some other matters can be observed in looking at Figure 5. Notice, for example, that the distance (on the logit scale) between the thresholds is quite variable, with some distances (within a single) item being up to twice the distance of others. For example, for item C1, note that the distance between the 4th and 5th thresholds is over 3 times the distance between the 1st and 2nd. This result makes it clear that the Likert assumption that the raw-score responses are at an interval level is not supported. A similar finding was reported by Embretson (1996), who saw it as a general result that pertained quite generally across many contexts. Moreover, beyond that, it is also clear that the items themselves (essentially the item stems) also vary in their difficulty, with item C1 being about 2 logits easier to agree with than item C5—translating that into probabilities, that is a difference of about .25 (or 25%)—that is, if a student's probability of agreeing with item C1 were 0.50, then their probability of agreeing with item C5 would be predicted to be 0.25, a very considerable difference. This sort of result, which is also quite common in such circumstances, means that the interpretations of the summated scores from Likert scales will depend on which items are chosen to include in the instrument—instruments with a fixed set of items will not be affected by this, but usage of item banks, or even just different sets of items, as in, say a pretest-posttest design, will be complicated in such cases. Note that the concerns expressed in this paragraph relate to limitations in interpretation based on the raw scores (i.e., summated scores) for the instrument, which is quite common in psychological applications. When translated into the context of a modern psychometric analysis, such as that used here, these concerns are ameliorated, if not eliminated altogether.

This pattern of diffuse and overlapping thresholds has been found to be very common in analyses of data from Likert response format items in our experience with many attitude scale developments at the BEAR Center. While we have been dismayed by this finding of the vagueness of the Likert options, we have not been moved to abandon the possibility of having a way to establish good evidence for internal structure validity for properties such as those that are usually measured using Likert response format items. Instead, we have been making attempts to develop an alternative means of transducing the attitudinal-like properties using the Guttman response format, so that the possibilities for establishing internal structure validity are increased.

The Guttman response format items.

The SFHI Project also administered 12 Guttman items to the same sample of students who took the Likert items at the same time. The separation reliability[5] of the Guttman item set was 0.87—this value compares well with reliability for the 45 Likert-style items, which were found to be slightly higher at 0.89—one way to think about this is to note that the

[5]The separation reliability is defined as the equivalent to Cronbachs' alpha, except that in calculating it, the logit estimates of the student estimates are used in place of the (traditional) raw scores (Wilson, 2005; Wright & Masters, 1982).

Likert-style item set has 120 (= 45 X 5) category-pair comparisons, while the Guttman-style item set has 48 (=12 X 4), less than half. The (directly-estimated[6] and hence disattenuated) correlation between the two sets of results (i.e., Likert and Guttman) was found to be 0.82.

The items set was examined with the same validity evidence criteria as for the Likert response format items reported above. Figure 6 shows the Guttman equivalent of the Likert Wright Map[7] (which was shown in Figure 5). The conventions for this Figure are the same as for the earlier Wright map. In this case, however, the pattern of the thresholds is more consistent. The project developers applied a standard-setting process called "construct modeling" (Draney & Wilson, 2011) to come up with bands that attempt to maximally separate the thresholds levels across the items. This has led to a much clearer delineation of the segments of the logit scale associated with each of the levels, as indicated by the horizontal bands shown in the Figure. Note that there is still some overlap between the RIS construct map bands here also, for example, note that for item 10, the 10.1 threshold has been judged just above the band for construct level 1. The exception here is Item 4, which has been harder for the students to agree with at each of the levels except the highest—this item should be re-considered for inclusion in the final set of Guttman items. Hence, for the Guttman response format items, the internal structure validity evidence is much clearer —indeed, across the item set, with the possible exception of Item 4, the item responses have generally conformed to the predicted order as indicated in the RIS construct map.

Looking beyond the accumulation of internal structure validity evidence, one can see a further advantage that accrues to the Guttman-style scale—the establishment of the bands provides a criterion-referenced interpretation of values on the scale. For example, a student estimated to be at 1.0 logits would be interpreted as one whose researcher identity can be said to be at the level where they are "comfortable" (see Figure 2) with their researcher identity, though still at the lower end of that category. For a student at the location 1.0 on the Likert-style scale looking back at Figure 5), all one could say is that their responses would range from "slightly agree" to "agree" on the items, which is a less interpretable, and therefore a much less meaningful statement.

Comparing the Guttman response format scale with the Likert response format scale.

The results discussed above show a considerable difference in interpretability between the Guttman-style scale with the Likert-style scale. The data produced by this study affords an opportunity for a much more direct and detailed comparison between the two, as the same sample of students responded to both forms, and there was a matching of at least some of the Likert-style items with the Guttman-style items. This latter opportunity arose from the way that the two item sets were developed, as described above.

[6]A two-dimensional PCM estimation was used for this.
[7]Note that, even though this is based on the same sample of students, the actual logit values on the two Wright maps cannot be directly compared as there are no common (or link) items between these two analyses.

A third analysis was conducted to compare performance of the matching 20 Likert response items and the 12 Guttman response items. This was, again, a unidimensional PCM analysis, and, as the two sets of items are both included, the logit values here can be directly compared between Likert response format estimates and Guttman results. The item difficulties for the Guttman response format items were anchored from the previous analysis, linking the scale in Figure 6 with this one, and the student estimates were generated, based only on the matching Likert response format items. The Wright Map showing the results for this analysis is given in Figure 7. The banding established using the Guttman response format items has been applied to this Wright map (in Figure 7), so now we can see how the student estimates from the Likert response items are related to the levels of the RIS construct map (as shown in Figure 2). In fact, the two distributions, one based on the Guttman response format items (in Figure 6) and the one based on the Likert response format items are very similar—which seems quite reasonable, given that most of the items were generated to match, and that the correlation between the two empirical variables was found to be 0.82 (as noted above). Thus, the measurement results for students who respond to the RIS Likert response format items can now be interpreted just as for those who took the Guttman response format items. For example, that student who was estimated to be at 1.0 logits on the Likert version, in contrast to the earlier interpretation (as noted at the end of the previous section) can now be interpreted as one whose researcher identity is at the level where they are “comfortable” (see Figure 2) with their researcher identity.

Discussion and Conclusion

The arguments presented in this paper have focused attention on certain limitations inherent in Likert response format items. It has been pointed out that interpretation of Likert’s original raw-score implementation has been questioned in the literature given that the raw scores do not have an interval nature. While this is a valid criticism of use of the unscaled raw-scores, we see this criticism as being largely overcome by applying appropriate latent variable scaling procedures, such as the Rasch-type models used in this paper. An observation that may help one to see why this use of raw sum-scores can still often be reasonably successful is to examine the test characteristic curve. For example, for the analysis reported in Figure 5 (i.e., for the Likert response format items), the test characteristic curve is shown in Figure 8. Examining this Figure, one can see that the relationship between the raw scores and the underlying latent variable (θ) is close to linear over much of the range of the raw scores, and hence, any statistical uses or manipulations that are not sensitive to translations of the underlying variable, such as those involving correlations, will not be affected by the use of raw sum-scores (except at the extremes of the scores). This includes classical factor analyses and many other common statistical procedures. The circumstances where this would indeed cause problems, or at least complications, would be situations that involved the flexible use of alternate subsets of items, such as in an item-bank situation, or situations where specific values on the underlying metric were important, such as the use of the bands in Figure 5 for interpreting the outcomes.

However, the use of latent variable scaling does not avoid every limitation of the Likert response format item form. As was shown in Figure 3, the split nature of the Likert response

format item, into a stem and standard options, makes it difficult to relate the empirical results with content structure such as is conveyed in a construct map. While the content of the stem could likely be coordinated with construct map levels, the uniformity of the option will generally count against that, even though they are an aspect of the efficiency that Likert thought he had but was mistaken. Moreover, the assumption that the standardized content of the options will result in equal interval raw scores for the items is seen to be an aspiration, but, in most cases, as here for the RIS scale, this will not be a sound assumption. This causes problems for both the possibility of interval-level outcomes, but also for achieving some sort of consistent match with levels in a construct map. Hence, there is a gap in the possible validity evidence for instruments using Likert response format items.

Turning now to discuss aspects of the Guttman response format item, one can find in the literature several places where experts have criticized the use of Guttman Scales (scalogram, etc.) as being less than desirable. These include De Vellis (2017), and Nunnally (1967) the latter of whom, while complimenting the intuitive nature of Guttman scaling criticized its “fundamental impracticality.” However, these criticisms, and others like them, are criticisms of the practice of Guttman scaling, not of the Guttman-formatted item including its gradational response options, as described here, and hence these criticisms do not apply to this work.

The idea of the Guttman response format item was developed directly as a way to build content validity into the item, and thus, to allow one to examine support for that in terms of internal structure validity—specifically by examining the resulting item parameter estimates for consistency with the levels of the construct map. The pattern of results shown in Figure 6 displays a strong consistency between the empirical results and the intended structure of the RIS construct. There is not 100% consistency, but then we expect some imperfections in our item realizations—it is certainly a considerable improvement over the results shown in Figure 5. This addition to core validity evidence (which is so often lacking in attitude scale and survey development) should be seen as a very important contribution of the Guttman response format item.

This advantage comes with no great cost in terms of net reliability of the instruments, which was found to be only slightly different between the two item formats, 0.89 for the Likert response format item set and 0.87 for the Guttman response item set. Of course, this net comparison might be misleading, as the number of items is hugely different between the two (45 versus 12), and the number of categories within each item are also different (6 versus 5, or, more directly, 5 versus 4 ordered distinctions)[8]. Given the near equality of the reliabilities, one can say that, for this specific pair of Likert and Guttman items, each Guttman response format item is worth (in terms of contribution to reliability)

[8]One can use the Spearman-Brown formula to make these approximately equivalent by transforming the problem to one analogous to a dichotomous item situation. Using the informal rule that a polytomous item with k categories is equivalent to approximately k-1 dichotomous items (Donohue, 1993), this means that the Likert set may be seen as being like a set of 125 dichotomous items, while the Guttman set may be seen as the equivalent of 48 dichotomous items. Thus, one can see that, if one had the same number of distinctions available in the Guttman set as were available in the Likert set (125), then one would have 125/4 = 31 (approx.) Guttman items, and applying the Spearman-Brown formula, this would give an equivalent reliability of a test that is 125/48 = 2.6 times as long (in dichotomous item equivalents): (2.6X0.87)/(1.0 +0.87) =1.21, or, given that 1.00 is the maximum of reliability, 1.00.

approximately (45/12=) 3.75 Likert-style items. This is a considerable efficiency and should be seen as a second important contribution.

However, there is a caveat to this second contribution. The work to develop Guttman-response format items can readily be seen to be more content-intensive. That is, where a typical Likert-style item will need the creation of only a single stem statement, a typical Guttman-style item will need four or five such statements (e.g., one for each construct map level). Hence, the item development process will likely be more time-consuming than for Likert response format items in terms of items developed per unit of work-time—this has previously been reported in an early example of such work (Teh, 2003). Secondly, we have found that, when respondents respond to Guttman response format items, they often report that their rate of response is slower than for items formatted in the Likert response style—and this is consistent with the development demands, as the respondents need to read more lines of text for the Guttman-style items. Thus, although the Guttman response format items are seen to have a considerable advantage over Likert response format items in terms of contribution to reliability, this will need to be balanced against a slower rate of item development and a slower rate of response.

In sum, one can evaluate the potential of the Guttman response format item that has been the focus of this paper as being most useful for (a) focusing the work of those who are developing instruments in the social sciences on founding their development in a content-based model of the construct (such as the construct map, as used here), (b) helping establish the construct validity (in terms of internal construct validity) for the instrument, (c) adding to the meaningfulness of the outcomes that get reported, and (d) providing a path for establishing internal structure validity evidence for the numerous Likert Scales that are used by social scientists today.

Acknowledgments

This project was supported by the National Institute of General Medical Sciences, the National Institutes of Health, under Award Number 8R25GM129194. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix A

The left column of Figure A1 shows the 21 chosen Likert-style items. Each of these Likert-style items was developed with six ordered response choices – strongly agree, agree, slightly agree, slightly disagree, disagree and strongly disagree. To transform the set into to the Guttman-style format, we first grouped Likert-style items together based on similar content. For example, the first three Likert-style items in Figure A1 are focusing on an individuals' comfort with seeing himself/herself as a researcher. The first item "I am beginning to consider myself a researcher," targets a relatively lower level of the construct map in comparison to the second item "I consider myself a researcher," which in turn targets a relatively lower level of the construct map in comparison to the third item "I consider myself to be a professional researcher." Each of these items becomes an option in the first Guttman-style item (G1), adapted in some cases to make the expressions consistent across the Guttman-style options. To match to the construct map for this variable, we added two

more options at levels below that for Item 1. The development process was similar for four of the Guttman-style items. For some Likert-style items (e.g., Item 7, and five others) there were no others that were of a similar content, so we needed to add four Guttman-style options for each of them. We found that this process produced a somewhat imbalanced set of Guttman-style items, with two Guttman-style items for the Agency strand and three for the rest. Hence, we developed one extra Guttman-style item focused on the Agency strand that was not matched among the 21 Likert-style items (Guttman-style item G9). Through this process, we obtained the right column of the Figure, the 12 Guttman-style items. In subsequent analyses, one of the Likert-style items (Likert-style item 12) was found not to fit the statistical model, so we deleted it from the comparisons, but the corresponding Guttman-style item did fit, so we left it in the comparisons. Each of the remaining 20 Likert-style items maps to an option for one of the Guttman-style items. To get comparable estimates, we ensured each student in our sample of 863 high school students took both formats of the instrument. We a randomized the order of the instruments, meaning, some students looked at the Likert items first while others looked at the Guttman items first.

Likert-Style Items	Guttman-Style Items
	G1. Which statement about being a researcher best captures your opinion of yourself? a) I do not consider myself a researcher. b) I probably do not consider myself a researcher.
1. I am beginning to consider myself a researcher.	c) I am beginning to consider myself a student researcher.
2. I consider myself a researcher.	d) I consider myself a student researcher.
3. I consider myself to be a professional researcher.	e) I consider myself to be a professional researcher.
	G2. Which statement below best describes your skills to do research? a) I do not have the skills to do research. b) I'm interested in gaining research skills.
4. I have the basic skills to help do research.	c) I have the basic skills to help do research.
5. I have the skills to help do research.	d) I have the skills to conduct research with a little help from others.
6. I have the skills to conduct research on my own.	e) I have the skills to conduct research on my own.
	G3. Which statement below best captures your identity as a researcher? a) Being a researcher is not a part of who I am. b) I am not sure if being a researcher is a part of my identity. c) Being a researcher might be a small part of my identity. d) Being a researcher is a part of my identity.
7. The researcher part of my identity is important to me.	e) Being a researcher is a big part of my identity.
	G4. Which statement best describes you? a) I don't consider myself a part of a research community. b) I am beginning to feel like a part of a research community.
8. I am a member of a research community	c) I am a small part of a research community.
9. I am a part of a group of researchers.	d) I am a part of a research community.
10. I am an important part of a group of researchers	e) I am an important part of a research community.
	G5. Which statement best describes your interest in research? a) I do not have an interest in doing research that helps my community. b) I am slightly interested in doing research that helps my community. c) I might be interested in doing research that helps my community.

11. A career in research would be a good way for me to help people.	d) I would be interested in doing research that helps my community. e) I am definitely interested in doing research that helps my community.
12. I am comfortable talking with more experienced researchers.	G6. Which statement best describes your level of comfort in communicating with researchers? a) I am uncomfortable speaking to experienced researchers right now. b) I hesitate to speak to researchers that have more experience than me. c) I am learning how to communicate with researchers that have more experience than me. d) I am comfortable talking to researchers that have more experience than me. e) I can speak with confidence to researchers that have more experience than me.
13.I have a strong desire to make a meaningful contribution to society through research.	G7. Which statement best describes your interest in contributing to society? a) I do not have the desire to contribute to the society through research. b) I have an interest in contributing to society through research. c) I have a desire to make some contribution to the society through research. d) I have a desire to make a meaningful contribution to the society through research. e) I have a strong desire to make a meaningful contribution to the society through research.
14. I can research issues independently.	G8. Which statement best describes your level of skill? a) I have no research skills. c) I can research issues with a lot of help c) I can research issues with some help d) I can research issues with very little help e) I can research issues independently.
	G9. Which statement best describes your level of researcher voice? We describe researcher voice as the extent to which you feel empowered to speak about your research. a) I do not want to have a researcher voice. b) I do not have a researcher voice now but I would like to develop one. c) I can use my researcher voice to guide discussions. d) I am developing a strong researcher voice. e) I have a strong researcher voice.
	G10. Which statement best describes your future plans? a) I do not plan to pursue research in the future. b) I do not know if doing research is in my future.

	c) I am not sure if a research-related degree is right for me. d) I might get a research-related degree in college.
15. I plan to get a research-related degree in college.	e) I plan to get a research-related degree in college.
	G11. Which statement best describes your interest in research?
	a) I think research is boring.
	b) I think research is a little interesting.
16. I think research is interesting.	c) I think research can be interesting at times.
17. I think that research is an exciting field of study.	d) I think research is very interesting.
18. I think research is an engaging field of study.	e) I think research is an engaging field of study.
	G12. Which statement best describes your interest in a research career?
	a) A career in research would not be a good fit for me.
	b) I am not sure if I am interested in research as a career.
19. I think a career in research maybe a good fit for me.	c) I might have an interest in research as a career.
20. I think a career in research could fit into my career plans.	d) A career in research could be a good fit for me.
21. A career in research would be a great choice for me.	e) A career in research would be a great fit for me.

Figure A1.
Mapping Likert-style Items into Guttman-style Items.

References

Adams RJ, Wu ML, Cloney D, & Wilson MR (2020). ACER ConQuest:Generalised Item Response Modelling Software (Version 5) [Computer software]. Camberwell, Australia: Australian Council for Educational Research.

AERA, APA, NCME (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Bathia S, Morell L, Wilson M, Koo B, & Smith R, (2020). Exploring high school student's self identification with scientific research. BEAR Center Research Report, University of California, Berkeley.

Carifio J, & Perla RJ, (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert Scales and Likert response formats and their antidotes. Journal of Social Sciences, 3(3), 106–116.

DeVellis RF, (2017). Scale Development: Theory and Applications 4th edition. Sage Publications, Inc.

Donohue JR, (1993). An empirical examination of the IRT information in polytomously scored reading items. Princeton, NJ: ETS Research Report, RR-93-12.

Draney K, & Wilson M, (2011). Understanding Rasch measurement: Selecting cut scores with a composite of item types: The Construct Mapping procedure. Journal of Applied Measurement, 12(3), 298–309. [PubMed: 22357129]

Embretson SE, (1996). The new rules of measurement. Psychological Assessment, 8(4), 341–349.

Glass GV, Peckham PD, & Sanders JR, (1972). Consequences of failure to meet assumptions underlying the analyses of variance and covariance. Review of Educational Research, 42, 237–288.

Guttman L, (1944). A basis for scaling qualitative data. American Sociological Review, 9, 139–150.

Guttman L, (1950). The basis for scalogram analysis. In, Stouffer SA, Guttman L, Suchman EA, Lazarsfeld PF, Star SA, Clausen JA (Eds.), Measurement and Prediction (pp. 60–90). Princeton, NJ: Princeton University Press.

Hambleton RK, Swaminathan H, & Rogers HJ, (1991). Fundamentals of Item Response Theory. Newbury Park, CA: Sage.

Jamieson S, (2005). Likert scales: How to (ab)use them. Medical Education, 38, 1212–1218.

Kofsky E, (1966). A scalogram study of classificatory development. Child Development, 37, 191–204.

Koo BW, Bathia S, Morell L, Gochyyev P, Phillips M, Wilson M, & Smith R, (2021). Examining the effects of a peer- learning research community on the development of students' researcher identity, confidence, and STEM interest and engagement. Journal of STEM Outreach, 4(1).

Kuzon WM Jr., Urbanchek MG, & McCabe S, (1996). The seven deadly sins of statistical analysis. Annals of Plastic Surgery, 37, 265–272. [PubMed: 8883724]

Labovitz S, (1967). Some observations on measurement and statistics. Social Forces. 46(2), 151–160.

Morell L, Bathia S, Koo B, Gochyyev P, Wilson M, & Smith R, (April 10, 2021). A Survey to Measure Secondary School Students' Identity in Research (IR-HS). In session, *Identity as an Outcome and Unfolding Process in Science Education*. NARST Annual Meeting Orlando, FL (Remote Conference).

Nunnally JC, (1967). Psychometric Theory. New York: McGraw-Hill.

Osgood CE, & Tannenbaum PH, (1955). The principle of congruence in the prediction of attitude change. Psychological Bulletin, 62, 42–55.

Rasch G, (1960). Probabilistic Models for Some Intelligence and Attainment Tests. [Reprinted by University of Chicago Press, 1980].

Robinson MA, (2018). Using multi-item psychometric scales for research and practice in human resource management. Human Resource Management, 57, 739–750

Schwartz R, Ayers E, & Wilson M, (2017). Mapping a learning progression using unidimensional and multidimensional item response models. Journal of Applied Measurement, 18(3), 268–298 [PubMed: 29579739]

Teh LW, (2004). Development of an instrument to measure Singaporean student's attitude towards National Education Program. BEAR Center Research Report, University of California, Berkeley.

Thurstone LL, (1928). Attitudes can be measured. American Journal of Sociology, 33, 529–54.

Traylor M, (October 1983). Ordinal and interval scaling. Journal of the Market Research Society. 25(4), 297–303.

Uebersax JS, (2006). Likert scales: dispelling the confusion. Statistical Methods for Rater Agreement website. 2006. Available at: http://john-uebersax.com/stat/likert.htm. Accessed: 12/29/20.

Wilson M, (2005). Constructing measures: An item response modeling approach. Mahwah, NJ: Erlbaum (now published by Taylor and Francis, New York)

Wright BD, & Masters GN, (1982). Rating Scale Analysis: Rasch Measurement. Chicago: MESA Press.

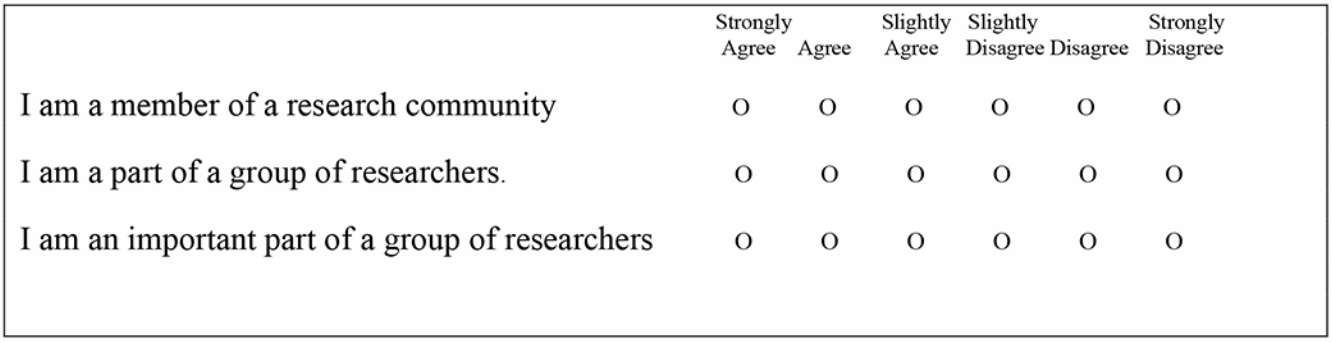

Figure 1.
Some example items from the Researcher Identity Scale (Likert response format items).

Construct Definition	Sample Items
Level 4: *Integration of Identity* Student identifies as a researcher and integrates this into their larger self	Agency - I can do research that benefits people. Fit & Aspiration - I plan to get a research-related degree in college.
Level 3: *Comfortable with Identity* Student begins to feel comfortable with their identity as a researcher	Agency - I can discuss research ideas with my peers. Self - I am beginning to consider myself a researcher.
Level 2: *Role Exploration* Student explores the different aspects of research	Community - I am making a contribution to a research group. Fit & Aspiration - I would like to do research.
Level 1: *Curious Identity* Student is a newcomer to the concept of research	Community - I am a member of a research community. Self - I can do research tasks with help from others.
Level 0: Absent Student is unaware of what research entails and has not considered their own role in research.	Self – I think research is boring.

Figure 2.
The construct map and some sample items from the RIS. (Note, adapted from Figure 1 in Bathia et al, 2020.)

5	If you were offered a good job, what would you do?
(a)	I would take the job
(b)	I would turn it down if the government would help me to go to school
(c)	I would turn it down and go back to school regardless
6	If you were offered some kind of job, but not a good one, what would you do?
(a)	I would take the job
(b)	I would turn it down if the government would help me to go to school
(c)	I would turn it down and go back to school regardless
7	If you could get no job at all, what would you do?
(a)	I would not go back to school
(b)	If the government would aid me, I would go back to school
(c)	I would go back to school even without government aid
8	If you could do what you like after the war is over, would you go back to school?
(a)	Yes
(b)	No

Figure 3.
Four of Guttman's items (1944).

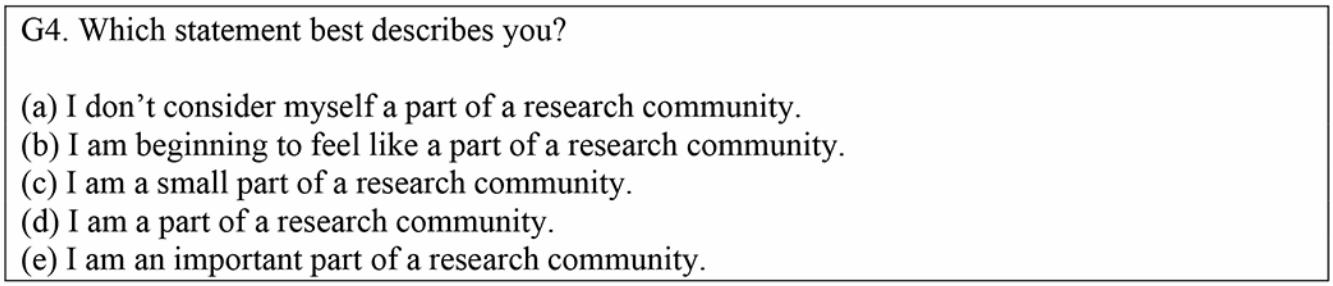

Figure 4.
An example item from the RIS (Guttman response format items).

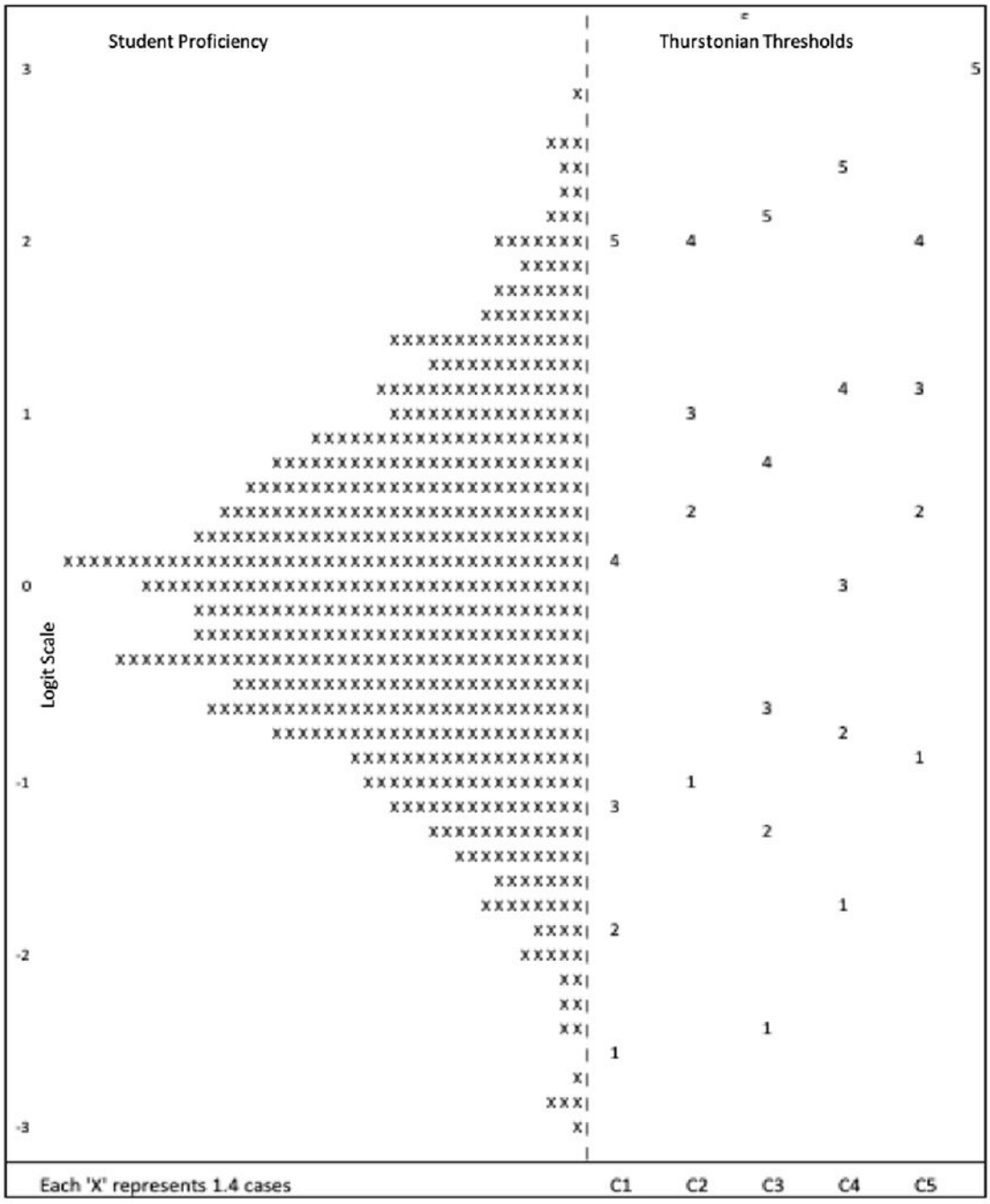

Figure 5.
The Wright map for the Community section of the RIS (Likert response format items).

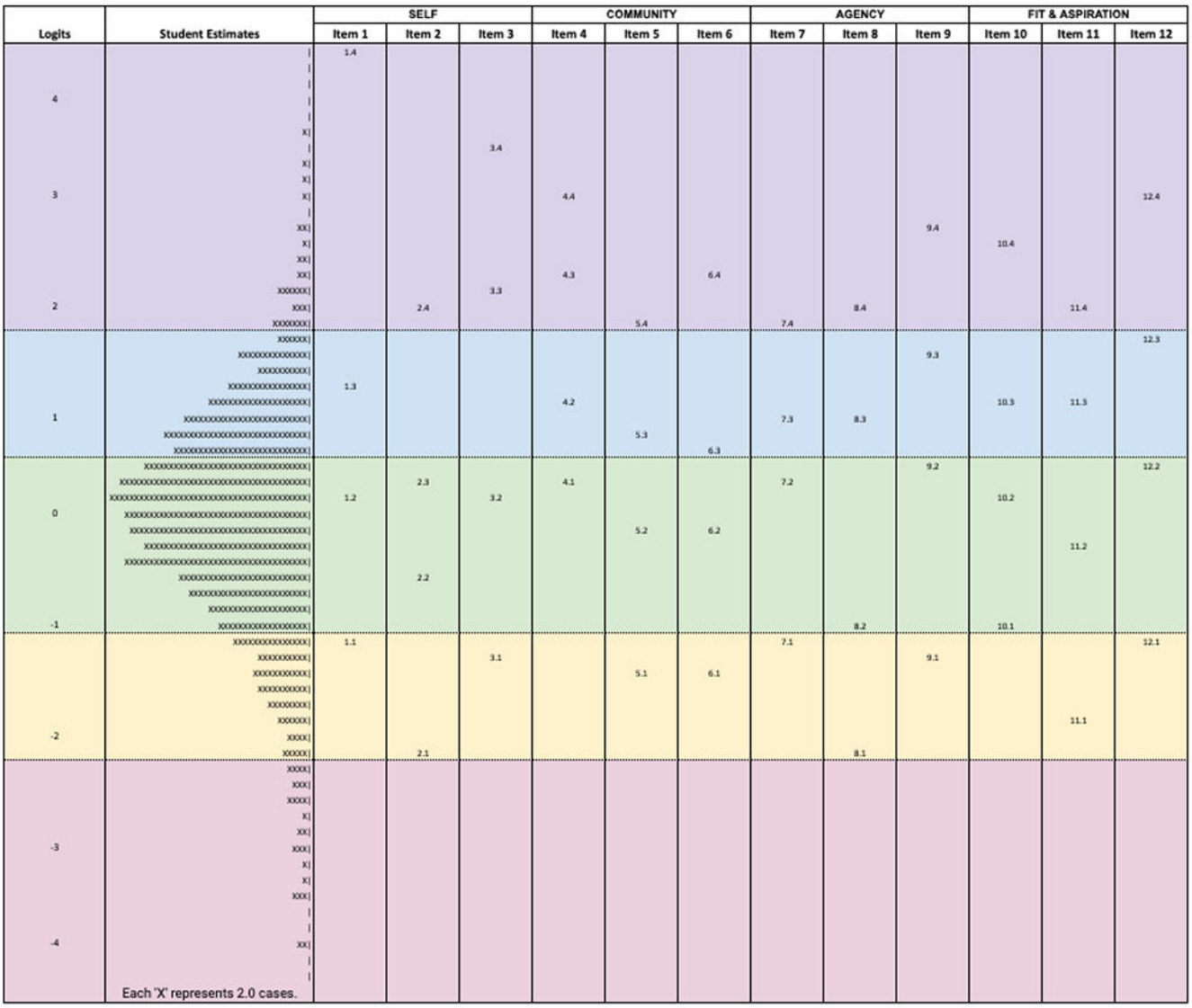

Figure 6.
The Wright map for the RIS (Guttman response format items).

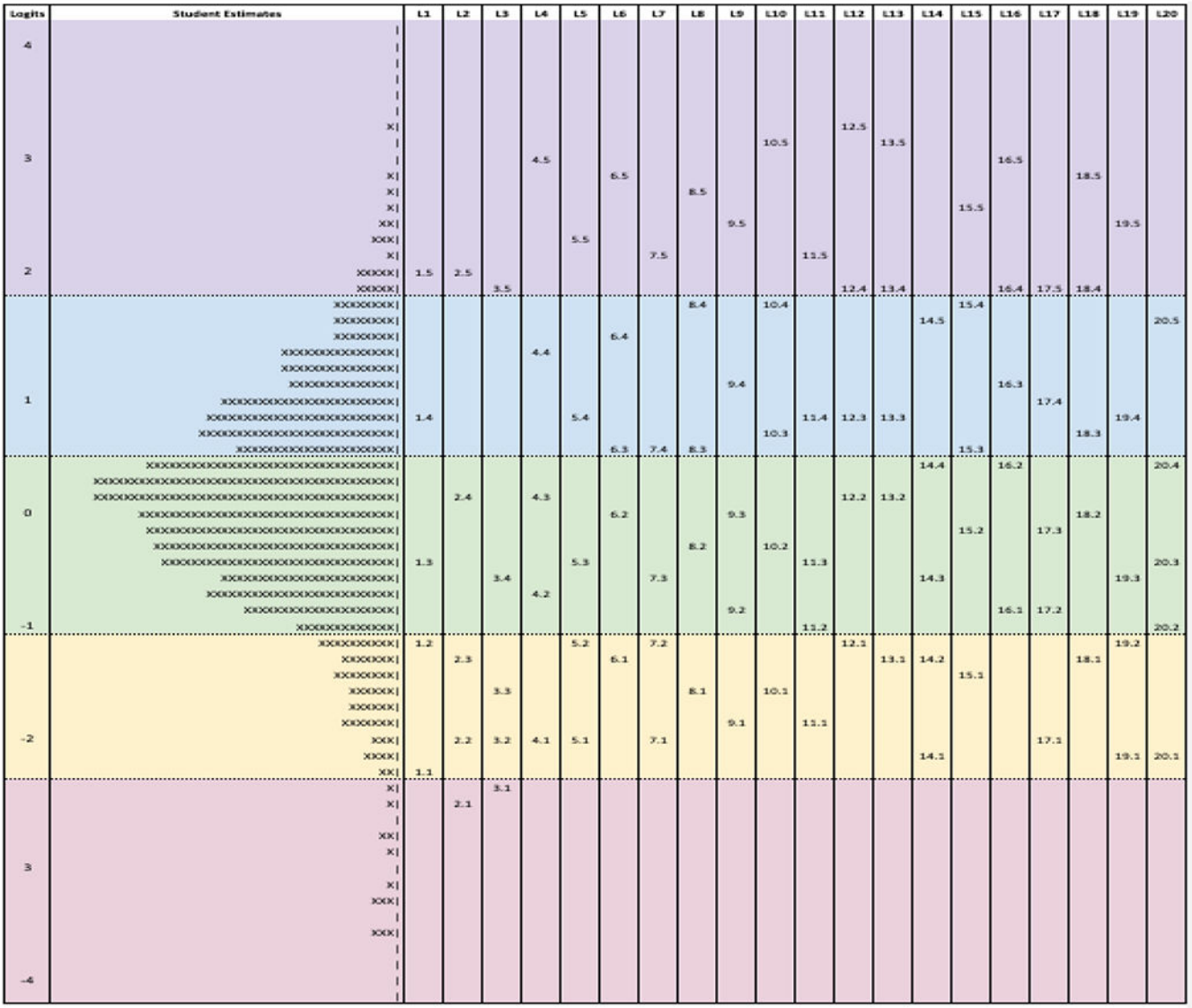

Figure 7.
The Wright map for the RIS (matching Likert response format items).

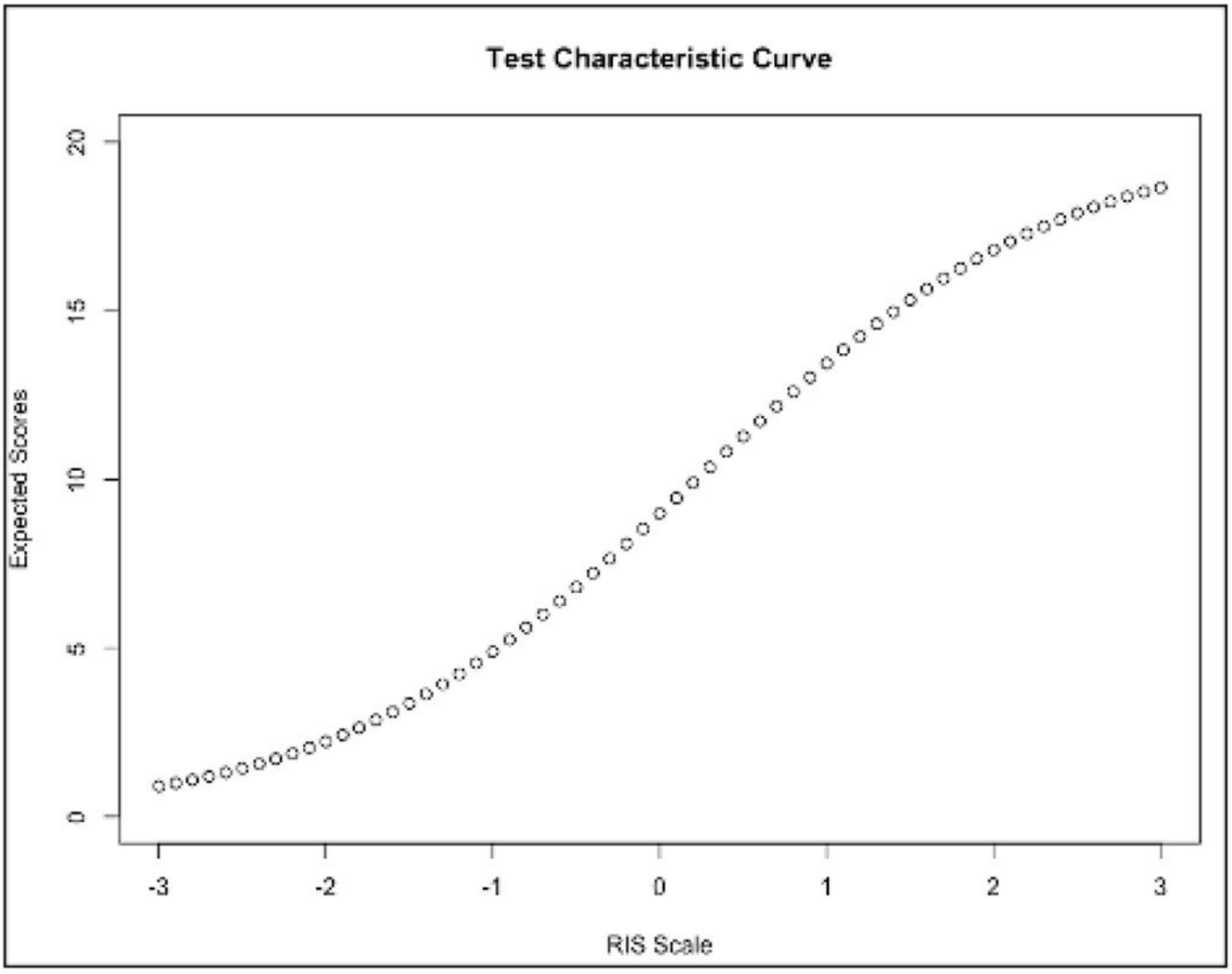

Figure 8.
The test characteristic curve for the Likert response format items.