

# The Relationship Between the Phi Coefficient and the Unidimensionality Index $H$ : Improving Psychological Scaling From the Ground Up

Johannes Titz  
Chemnitz University of Technology


To study the dimensional structure of psychological phenomena, a precise definition of unidimensionality is essential. Most definitions of unidimensionality rely on factor analysis. However, the reliability of factor analysis depends on the input data, which primarily consists of Pearson correlations. A significant issue with Pearson correlations is that they are almost guaranteed to underestimate unidimensionality, rendering them unsuitable for evaluating the unidimensionality of a scale. This paper formally demonstrates that the simple unidimensionality index  $H$  is always at least as high as, or higher than, the Pearson correlation for dichotomous and polytomous items ( $\phi$ ). Leveraging this inequality, a case is presented where five dichotomous items are perfectly unidimensional, yet factor analysis based on  $\phi$  incorrectly suggests a two-dimensional solution. To illustrate that this issue extends beyond theoretical scenarios, an analysis of real data from a statistics exam ( $N = 133$ ) is conducted, revealing the same problem. An in-depth analysis of the exam data shows that violations of unidimensionality are systematic and should not be dismissed as mere noise. Inconsistent answering patterns can indicate whether a participant blundered, cheated, or has conceptual misunderstandings, information typically overlooked by traditional scaling procedures based on correlations. The conclusion is that psychologists should consider unidimensionality not as a peripheral concern but as the foundation for any serious scaling attempt. The index  $H$  could play a crucial role in establishing this foundation.

**Keywords:** unidimensionality, Phi, Pearson correlation, Guttman scaling, factor analysis

Measurement implies that one characteristic at a time is being quantified. (McNemar, 1946, p. 298)

Claiming that a variable is measurable implies that it is

---

Correspondence concerning this article should be addressed to  Johannes Titz, Department of Psychology, Research Methods and Evaluation in Psychology, Chemnitz University of Technology, Wilhelm-Raabe-Str. 43, 09120 Chemnitz, Germany. E-mail: [johannes.titz@psychologie.tu-chemnitz.de](mailto:johannes.titz@psychologie.tu-chemnitz.de).

The author sincerely thanks Friederike Brockhaus, Matthias Hörr, Vivien Lungwitz, and Peter Sedlmeier for their constructive feedback on an earlier version of the article.

The ideas presented herein have not been disseminated previously. No funding to declare. No conflicts or competing interests to declare and no funding to disclose. This is mainly a theoretical and methodological contribution so there is no ethics approval/consent to participate/consent for publication. Note that a data set is analyzed, but this data comes from an archive. The data and analyses scripts used in this paper are publicly available on github: <https://github.com/johannes-titz/unidim>

©American Psychological Association, 2025. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. The final article is available, upon publication, at: <https://doi.org/10.1037/met0000736>

unidimensional (e.g., Falissard, 1999; Hattie, 1984; Ziegler & Hagemann, 2015). This is evident for non-psychological variables. For instance, length, mass, and speed are unidimensional variables with fixed units. Two lengths expressed in the same unit can be added, which proves to be extremely useful. However, it does not make sense to add length to mass, as it involves combining two different dimensions. This principle applies equally to psychological variables. For instance, calculating a sum score (or a factor score) for *intelligence* from several items implies that the items measure the same dimension. When testing a hypothesis regarding *intelligence*, the sum score representing *intelligence* must not be a mixture of *intelligence*, *attention* and *concentration*. If it were, any failed hypothesis could be attributed to measurement problems (e.g., Ziegler & Hagemann, 2015). Therefore, before testing a hypothesis, it is crucial to establish beyond reasonable doubt that the variables in question are scalable and thus unidimensional.

There is a straightforward visual analogy: if a variable is scalable, one can draw a line and position objects (e.g., items and participants) on this line. The *single* line represents the concept of unidimensionality. This idea is simple and makes unidimensionality a highly intuitive concept. Despite this, psychologists have made unidimensionality seem complicated. Often, sophisticated factor-analytic methods with unrealistic assumptions and arbitrary cutoff criteria are used to test if

several items are unidimensional. Many psychologists are dissatisfied with these methods (e.g., Falissard, 1999; Hattie, 1985; Slocum-Gori & Zumbo, 2011; Ziegler & Hagemann, 2015) but are hesitant to adopt alternatives. For instance, Guttman scaling (Guttman, 1944), which exclusively focuses on unidimensionality, is rarely used in psychological research. Consider the European publisher Hogrefe, which offers approximately 700 psychological tests on their German website, with only one of these tests mentioning Guttman scaling in its detailed description.

One reason for this neglect might be that psychologists do not believe Guttman scaling offers a fundamentally different perspective. Most psychologists prioritize inter-item correlations; if these are high, both factor analysis and Guttman scaling suggest one dimension. Conversely, if the correlations are low, neither method strongly supports a single dimension. However, this assumption is flawed. A scale can be perfectly unidimensional even with medium or low item correlations. Factor analysis can incorrectly suggest multiple dimensions when there is only one. This article demonstrates when and why this discrepancy occurs and explores the implications for psychological scaling.

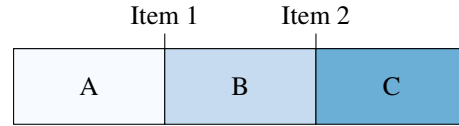
In the first part, perfect scaling (Guttman scaling) is introduced as a conceptualization of unidimensionality. Then, the relationship between the unidimensionality index  $H$  and the Pearson correlation for dichotomous and polytomous items ( $\phi$ ) is examined. Using a derived inequality, a theoretical scenario is presented where factor analysis fails to detect a single dimension, even though the data is perfectly unidimensional. In the second part, an empirical investigation demonstrates that this issue can also occur with real data. Finally, the implications of these findings for psychological scaling are discussed.

### Part I: Theory

Many psychologists are not familiar with Guttman's scaling ideas, so a motivating example is presented. Imagine an intelligence test that consists of two dichotomous items measuring the same underlying latent variable. One item is easy, and the other is difficult. You want to assign a single meaningful value to each participant regarding the underlying latent variable (intelligence). If the process is deterministic and all confounding variables are known, there can only be three outcomes: (1) both items are not solved, (2) both items are solved, or (3) the easy item is solved, but the difficult one is not (for a similar example, see Guttman, 1944, p. 143). It is not possible for the difficult item to be solved while the easy one is not, as this would be a logical contradiction. This becomes evident when considering a one-dimensional variable (a line) on which both items and all participants are located (see Figure 1).

With two items, four response patterns are possible, but only three regions exist on the line in Figure 1. Therefore, one

**Figure 1:** An example of perfect unidimensional scaling with two dichotomous items



*Note.* The items and persons are positioned on a single latent scale (line). The two items divide the scale into three regions: Region A, before Item 1, with a score of 0; Region B, between Item 1 and Item 2, with a score of 1; and Region C, beyond Item 2, with a score of 2.

response pattern cannot occur. In region A, no item is solved, the response pattern is 00, and the score is 0. In region B, the easier Item 1 is solved, the response pattern is 10, and the score is 1. In region C, both items are solved, the response pattern is 11, and the score is 2. The remaining response pattern, 01, is not allowed because it violates unidimensionality. A person cannot be located above Item 2 and simultaneously below Item 1 (pattern 01). This principle is the essence of unidimensional scaling. Simply counting how often the violating pattern occurs serves as a reasonable unidimensionality index<sup>1</sup>, with many more complex indices derived from it (for an overview see, for instance, Zysno, 1993).

Note that Guttman's approach is purely non-parametric, meaning it does not require any assumptions about the distribution of the latent variable. In fact, this type of scaling relies on very few assumptions overall, as it is ordinal in nature. The procedure involves ordering items and participants without making any conclusions about the distances between them. This raises the question of whether unidimensionality is inherently tied to ordinality or if Guttman's model defines unidimensionality too narrowly, necessitating additional conceptualizations. This discussion will be deferred, as it is linked to the analyses that will be presented.

An extension of Guttman scaling to items with more than two ordered categories (polytomous items) was first proposed by Molenaar (1982) and subsequently developed independently by Zysno (1993). An introductory text can be found in van Schuur (2011). The basic principle of scaling remains the same: positioning items and participants on a single line. Since the items have multiple answer options, it is necessary to position the item *steps* rather than the items themselves on the line. For two items with three ordered categories, one of several possible unidimensional scales is depicted in Figure 2.

The maximum possible score is 4, and the minimum is 0, resulting in 5 regions, corresponding to 2 steps per item (from 0 to 1 and from 1 to 2). Out of the 9 possible response patterns, 4 are not permitted. Similar to the dichotomous case, the number of these disallowed patterns serves as the basis for unidimensionality indices.

It is important to understand that such unidimensionality

<sup>1</sup>0 indicating perfect unidimensionality.

**Figure 2:** An example of perfect unidimensional scaling with two trichotomous items

	Step 1 <sub>1</sub>	Step 2 <sub>1</sub>	Step 2 <sub>2</sub>	Step 1 <sub>2</sub>
A				
B				
C				
D				
E				

*Note.* The item steps and persons are positioned on a single latent scale (line). The four item steps partition the scale into five regions. The number represents the item, while the subscript indicates the step.

indices are not measures of correlation and can therefore differ from them. While high correlations approaching 1 trivially suggest unidimensionality, the interpretation becomes less clear for correlations that are low or moderate. Similarly, one may question whether a high unidimensionality index necessarily indicates a high correlation.

To explore these questions, it is essential to provide formal definitions and a specific scenario. While the principles of Guttman scaling apply to both dichotomous and polytomous items, this article will primarily focus on dichotomous items. There are two main reasons for this focus: Firstly, reasoning about dichotomous items is simpler. Most psychologists are already familiar with a fourfold table and the  $\chi^2$ -test, enabling them to follow the arguments more easily. Secondly, the item response process for dichotomous items is more straightforward. For example, in most ability tests, questions can be scored as 0 (not solved) or 1 (solved). Although there may be some room for interpretation, the response process is generally clear. In contrast, when participants respond to personality questions using a 5-point scale with semantic labels, the response process becomes more ambiguous. Therefore, it is more feasible to find an empirical example of a unidimensional scale for dichotomous items, which will be attempted here. It is important to note that the main findings also apply to polytomous items, which will be demonstrated but not as extensively discussed.

The focus will be on the Phi-coefficient ( $\phi$ ) as a measure of association and the  $H$ -index (Loevinger, 1947, 1948) as a measure of unidimensionality. With over 70 binary measures available (e.g. Brusco et al., 2021; Choi et al., 2010), it is important to explain why Phi and  $H$  were chosen. Notably, although the focus is on dichotomous items, both Phi and  $H$  can be applied to polytomous items, allowing more generalizable conclusions.

The Phi-coefficient for dichotomous and polytomous items is equivalent to the Pearson correlation, one of the most widely used indices in psychology.<sup>2</sup> In a study on the base-rate influence on the similarity of binary measures, Brusco et al. (2021) describe the Phi-coefficient as a “popular exemplar” (p. 9) for the corresponding subset.

Regarding information content, the Phi-coefficient is considered by some scientists to be the most informative single

score for evaluating the quality of a binary classifier prediction, and it is strongly recommended in machine learning (Chicco, 2017).<sup>3</sup> There has even been a suggestion to replace the area under the curve of receiver operating characteristics (ROC AUC) with the Phi-coefficient for binary classification (Chicco & Jurman, 2023).

Studying  $H$  as a unidimensionality index is appealing for several reasons. The most straightforward definition of unidimensionality is based on Guttman scaling, which is non-parametric and does not require sophisticated statistical models. If there are no Guttman errors in a scale, it can always be constructed to be perfectly unidimensional.  $H$  measures the deviation from this ideal scale, whereas many other indices of unidimensionality are unrelated to Guttman errors.

Out of the many proposals for unidimensional indices,  $H$  is a more advanced measure because it accounts for the expected value under the null model (van Schuur, 2011; Zysno, 1993). Additionally, significance tests can be easily constructed for many different applications of  $H$  (van der Ark et al., 2008). In ordinal item response theory (Mokken analysis), the  $H$ -index is the de facto standard (Ark, 2012; Sijtsma & Molenaar, 2002; van Schuur, 2011). In the analysis of binary measures mentioned earlier (Brusco et al., 2021),  $H$  was categorized as *ungrouped*, indicating its unique features.

## Definitions

For the dichotomous case, the widely known cross table in Table 1 is useful for understanding the definitions.

**Table 1:** Cross table of two dichotomous variables.

		Item 2		
		0	1	
Item 1	0	a	b	a+b
	1	c	d	c+d
		a+c	b+d	a+c+b+d = N

The item difficulties  $p_1$  and  $p_2$  are defined as:

$$p_1 = \frac{(c + d)}{N} \quad (1)$$

$$p_2 = \frac{(b + d)}{N}$$

To determine the number of violations regarding unidimensionality, it is useful to know which item is more difficult. Let

<sup>2</sup>A Google Scholar search returns 815,000 documents with the search term “Pearson correlation” psychology. The search term “Phi-Coefficient” psychology yields 14,000 documents, “tetrachoric correlation” psychology results in 6,000, and the search term “polychoric correlation” psychology returns about 10,000 documents.

<sup>3</sup>In this field, it is known as the Matthews correlation coefficient (MCC).

us assume the row item is easier or both items have equal difficulty, which leads to  $b$  being smaller than  $c$  or equal to  $c$ :

$$\begin{aligned} p_1 &\geq p_2 \\ c &\geq b \end{aligned} \quad (2)$$

Since we can always place the easier item in the row, we do not need to consider the case where the column item is easier.

For the dichotomous case, the  $\phi$ -coefficient serves as a useful shortcut for the Pearson correlation:

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}} \quad (3)$$

It will suffice to study positive correlations, as a negative correlation between two items can be made positive by reverse coding one of the items.

To understand how correlations are related to unidimensionality, an index of unidimensionality is required. Among the many different indices, the most advanced ones account for the expected errors under a null model (e.g. van Schuur, 2011; Zysno, 1993):

$$u = 1 - \frac{e_{\text{obs}}}{e_{\text{exp}}} \quad (4)$$

Where  $e_{\text{obs}}$  are the observed errors and  $e_{\text{exp}}$  the expected errors. One prominent unidimensionality index is the coefficient of homogeneity  $H$ , which is based on the idea of statistical independence originally described by Loevinger (1947, 1948). A straightforward introduction, complete with numerous examples, is provided by van Schuur (2011).

When two items are independent, their joint probability is equal to the product of their individual probabilities. Given that we are dealing with absolute frequencies, this product must be multiplied by the sample size. This approach mirrors the logic of the  $\chi^2$  test, leading to the calculation of the error frequency as follows:

$$e_{\text{exp}} = (1 - p_1)p_2N \quad (5)$$

$$= \frac{a+b}{N} \frac{b+d}{N} N \quad (6)$$

$$= \frac{(a+b)(b+d)}{N} \quad (7)$$

The observed errors are simply  $b$  and the coefficient of homogeneity  $H$  can then be defined as follows:

$$e_{\text{obs}} = b \quad (8)$$

$$H = 1 - \frac{bN}{(a+b)(b+d)} \quad (9)$$

When examining the definitions of unidimensionality (Formula 9) and correlation (Formula 3) for dichotomous items,

it is evident that both are composed of the same variables (the frequencies  $a, b, c, d$ ). However, the relationship between these two coefficients, as well as the conditions under which they might be similar or different, is not immediately clear.

### The Relationship Between $H$ and $\Phi$

A particularly interesting and useful relationship between  $H$  and  $\phi$  (for positive values) is:

$$H \geq \phi \quad (10)$$

For readers interested in the details of this formalism, the Appendix provides two distinct derivations of the inequality along with supplementary notes. However, the subsequent arguments are generally comprehensible without requiring a deep dive into these derivations.

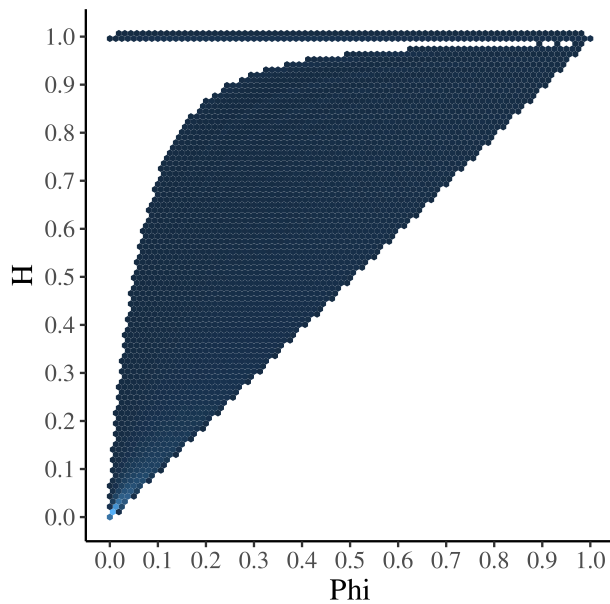
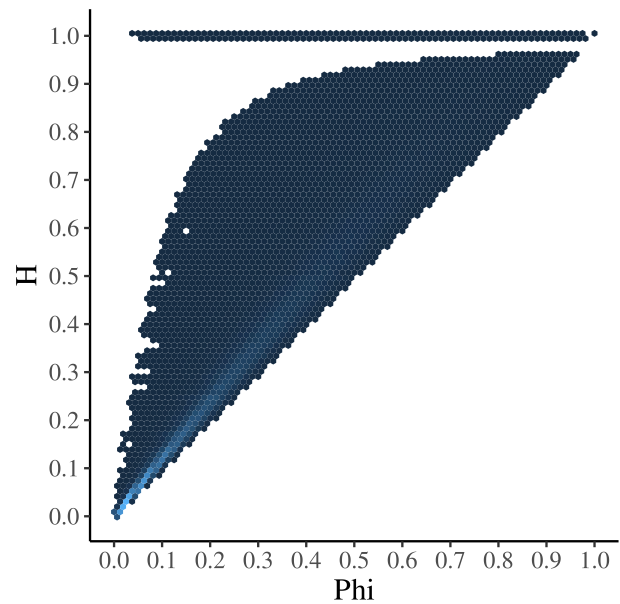
The inequality indicates that unidimensionality (as described by  $H$ ) is always at least as large as the Pearson correlation ( $\phi$ ). But how is this relevant to psychological scaling? The key insight is that  $H$  can be significantly larger than  $\phi$ , potentially reaching a value of 1. Pearson correlations are almost guaranteed to underestimate unidimensionality and are unsuitable for assessing the unidimensionality of a scale. This distinction is especially clear in the dichotomous case: under a perfectly unidimensional model with an error frequency  $b = 0$ , the correlation  $\phi$  remains unrestricted. For  $H = 1$ ,  $\phi$  can theoretically range anywhere between 0 and 1. This is because eliminating errors ( $b = 0$ ) does not impose meaningful constraints on  $\phi$ , illustrating the limitation of correlations in capturing unidimensionality.

$$\phi = \frac{ad}{\sqrt{a(a+c)d(d+c)}} = \frac{\sqrt{ad}}{\sqrt{(a+c)(c+d)}} \quad (11)$$

Setting  $c = 0$  leads to  $\phi = 1$ , while setting either  $a$  or  $d$  to 0 results in  $\phi = 0$ . If  $N$  is sufficiently large,  $\phi$  can take on any value in between, despite perfect unidimensionality.

To better illustrate this idea, the coefficient  $\phi$  was calculated for all combinations of  $a, b, c, d$  with  $N = 200$ , resulting in 1,373,701 combinations. Figure 3(a) shows the relationship between  $\phi$  and  $H$  for the 346,829 cases where  $\phi \geq 0$  and  $c \geq b$  ( $b$  is the error frequency). Additionally, a calculation was performed for a polytomous case with three ordered answer options. For this case,  $N = 30$ , resulting in 48,903,492 possible combinations, out of which 24,257,529 have a positive  $\phi$ . These results are depicted in Figure 3(b). Note that plotting such a large number of points is impractical, so Figure 3 instead shows hexagonal heat-maps, which aggregate the data based on regular hexagons.

Several observations are evident. First, when  $H = 1$ , all correlations can occur (evident from the line of hexagons at the top). This implies that even a very small correlation

**Figure 3:** *The relationship between  $H$  and  $\phi$* **(a)** *Two dichotomous items ( $N = 200$ )***(b)** *Two trichotomous items ( $N = 30$ )*

*Note.* Hexagonal heat-maps showing: (a) 346,829 cases where  $\phi \geq 0$  and  $c \geq b$  (the error frequency is  $b$ ) for  $N = 200$  with two dichotomous items (b) 24,257,529 cases where  $\phi \geq 0$  for  $N = 30$  with two trichotomous items. The lower limit is clearly visible in both subfigures, corresponding with  $H \geq \phi$ .

can be consistent with a unidimensional model. Second, the area between  $H = 1$  and the next smaller value of  $H$  is empty because the data is discrete and  $N$  is restricted. Certain combinations of  $H$  and  $\phi$  do not occur unless the sample size is large enough. The  $H$  values do not increase continuously; they can jump from lower values, such as .5 or .6, directly to 1 (for more details on this in the dichotomous case see Davenport & El-Sanhurry, 1991).<sup>4</sup> Third, and most importantly, the lower boundary as stated in Equation 10 is clearly visible. The higher the correlation, the higher the unidimensionality index. However, the unidimensionality index can be significantly larger, especially for small correlations. This is generally positive news, as factor-analytic results based on Pearson correlations may underestimate unidimensionality. Consequently, many questionnaires used by psychologists might be more unidimensional than currently believed. In extreme cases, factor analysis could completely overlook a perfectly unidimensional scale.

### When Factor Analysis Can Be Misleading

To illustrate how factor analysis based on  $\phi$  can be misleading, a unidimensional model ( $H = 1$ ) was created with 5 dichotomous items and 100 responses per item. The strategy to devise these items was straightforward. By setting  $b$  to 0 but  $c$  to a relatively large value, perfectly unidimensional items with low correlations were created. To avoid identical items,

$c$  was varied slightly.<sup>5</sup> All possible cross-tables of the 5 items are displayed in Table 2. In the header of the table (e.g.,  $1_0$ ,  $1_1$ , etc.), the first number is the item number, and the subscript number is the response (either 0 or 1). In all quadruples, the error frequency is 0 (depending on the difficulty, either  $b$  or  $c$ ). This data set is perfectly unidimensional. However, factor analysis based on  $\phi$  favors a 2-factor solution (see Table 3): the explained variance with one factor is only 53%, whereas with two factors, it is 90%. Although different criteria for factor extraction exist, to adequately recover the data, one must choose at least a 2-factor solution. Yet, the data is unidimensional.

Note that by using  $H$  instead of  $\phi$  as the input to factor analysis, one will always obtain the same or fewer dimensions because  $H \geq \phi$ . This holds true for both dichotomous and polytomous items. Clearly, fewer dimensions are desirable for many reasons, so what is the trade-off? Factor analysis based on  $\phi$  produces continuous factor scores, which can be conceived as an interval scale. In contrast,  $H$  essentially

<sup>4</sup>Since the calculation is quite computing-intensive, a much larger  $N$  is not feasible. However, this does not affect the argument presented here, which is based on the lower bound of  $H$ . Further note that for typical correlations in psychology of around .3, almost all  $H$  values above .3 are possible for the studied sample size.

<sup>5</sup>The source code to reproduce these results is available on GitHub: <https://github.com/johannes-titz/unidim>

**Table 2:** *Cross tables of five unidimensional items*

	1 <sub>0</sub>	1 <sub>1</sub>	2 <sub>0</sub>	2 <sub>1</sub>	3 <sub>0</sub>	3 <sub>1</sub>	4 <sub>0</sub>	4 <sub>1</sub>	5 <sub>0</sub>	5 <sub>1</sub>
1 <sub>0</sub>	90	0	10	80	89	1	11	79	85	5
1 <sub>1</sub>	0	10	0	10	0	10	0	10	0	10
2 <sub>0</sub>	10	0	10	0	10	0	10	0	10	0
2 <sub>1</sub>	80	10	0	90	79	11	1	89	75	15
3 <sub>0</sub>	89	0	10	79	89	0	11	78	85	4
3 <sub>1</sub>	1	10	0	11	0	11	0	11	0	11
4 <sub>0</sub>	11	0	10	1	11	0	11	0	11	0
4 <sub>1</sub>	79	10	0	89	78	11	0	89	74	15
5 <sub>0</sub>	85	0	10	75	85	0	11	74	85	0
5 <sub>1</sub>	5	10	0	15	4	11	0	15	0	15

**Table 3:** *Example of factor analysis for five unidimensional items*

Item	1-factor solution		2-factor solution	
	$\lambda$	$\lambda_1$	$\lambda_2$	
1	0.950	0.950	0.024	
2	0.119	0.093	0.946	
3	0.997	0.997	0.025	
4	0.125	0.099	0.993	
5	0.839	0.837	0.065	

indicates how well an ordinal model fits the data. In Guttman scaling, the equivalent of the factor score is the sum score, which is discrete and ordinal. Is the trade-off of sacrificing the interval scale worth it?

There does not seem to be a trade-off, as there is no compelling reason to argue that factor scores are on a meaningful interval scale in the first place. Factor analysis does not test for deviations from a true interval scale; it merely produces continuous values. There is no strong argument to suggest that the distance between factor scores of, for instance, 1.1, 1.2, and 1.3 are psychologically equivalent. To my knowledge, only one rigorous method exists to test whether latent constructs are on an interval scale: conjoint measurement theory (Luce & Tukey, 1964; Michell, 1990). So far, there is no evidence for interval scales in psychology, except perhaps for loudness and brightness perception (Luce & Steingrimsson, 2011). Thus, little seems to be lost when reverting to an ordinal model (for a similar argument see Heene, 2013).

Additionally, the question arises whether unidimensionality necessitates an interval or ratio scale, and if so, under which circumstances. Fundamentally, this poses the question of what unidimensionality truly is. Some authors are hesitant to provide a definitive answer. For instance, Heene et al. (2016) explicitly avoids a semantic definition and simply defines unidimensionality as local statistical independence. Ziegler

and Hagemann (2015) point out cases where this criterion is insufficient (see also Sijtsma & Molenaar, 2002). Sijtsma (2009) claims that unidimensionality is not a unitary concept but rather depends on the underlying model. While this pluralistic view is understandable, it overlooks the fact that unidimensionality is a fundamental concept of measurement. If it is not possible to clearly define such a fundamental concept, this might indicate overcomplication.

In the introduction, a general and simple definition of unidimensionality was given: if participants and items can be positioned on a single line, the construct could be regarded as unidimensional.<sup>6</sup> To be more specific, the participants and items must be positioned in such a way that the data can be fully recovered. This is the case for perfect Guttman scaling, where the sum score of a participant allows all item responses to be derived. Based on this perspective, unidimensionality does not require an interval or ratio scale, only order. By imposing unnecessary restrictions on tests of unidimensionality, such as the requirement for an interval scale, one reduces the likelihood of identifying unidimensionality when it exists. This could be summarized as the essence of  $H \geq \phi$ .  $\phi$  encompasses more than just unidimensionality.

Some critics might argue that the example presented in Table 2 is contrived, as it was explicitly chosen to show an extreme and unrealistic result (perfect unidimensionality). However, more realistic scenarios where  $H$  is smaller than 1 can also be easily created. Furthermore, there is no reason why the chosen response pattern could not occur empirically with real participants. Given the vast amount of data gathered in psychological research, it is very likely that similar cases do occur. This will be demonstrated in the second, empirical part of the paper.

## Part II: Empirical Example

Most currently used psychological questionnaires are probably not unidimensional because they are too complex. Consider the Big Five flagship NEO-PI-R (Costa & McCrae, 2008), which contains 240 items across 30 facets. For each of the five unidimensional factors, there are 48 items. It seems virtually impossible for 48 items to be strictly unidimensional.<sup>7</sup> The semantics and context of personality statements are too complex to consistently order 48 items on a single line.

For demonstration purposes, it is better to construct a new and simple example where unidimensionality has high

<sup>6</sup>Note that this is merely evidence, and the item response process could involve several dimensions behaving like one. This is a general problem of unidimensionality analysis: it is easier to provide evidence against a single dimension than to provide evidence for it (e.g. Dunn & Kalish, 2018). Still, as long as a phenomenon can be conceived as unidimensional, this should be favored until evidence is presented that shows otherwise.

<sup>7</sup>The NEO-PI-R uses polytomous items, but as has been shown previously, the inequality also holds for them.

face validity. A content area based on the author's expertise was chosen: teaching statistics. Consider the typical skills acquired by students studying research methods and statistics. For instance, to apply Bayesian revision, basic knowledge of probabilities is required. Without this basic knowledge, solving problems involving Bayesian revision is impossible. Such logical dependencies often lead to unidimensionality.

## Methods

The methods will only be briefly described, as this article is not a standalone empirical study. The aim is to present a real dataset that is almost perfectly unidimensional.

In a real statistics exam at the Institute of Psychology at Chemnitz University of Technology, 133 students were given the following scenario:<sup>8</sup>

You have been hired as an external consultant to evaluate the quality of a leadership assessment center for a large company. Based on theoretical considerations, you estimate that approximately 19% of all applicants are actually suitable for the position. Additionally, using data from numerous applicants and subsequent evaluations of their suitability, you can estimate two more probabilities: among the individuals who are actually suitable, 67% are rated as suitable by the assessment center; and among those who are actually unsuitable, 74% are rated as unsuitable by the assessment center. Your goal is to calculate the joint probabilities to subsequently estimate the conditional probabilities for incorrect decisions. What is the probability of the following events?

1. Probability of the conjunction of the two events: actually suitable AND rated as suitable by the assessment center (in %)
2. Probability of the conjunction of the two events: actually unsuitable AND rated as unsuitable by the assessment center (in %)
3. Probability of the conjunction of the two events: actually suitable AND rated as unsuitable by the assessment center (in %)
4. Probability of the conjunction of the two events: actually unsuitable AND rated as suitable by the assessment center (in %)
5. Probability that a person rated as suitable is actually suitable. (in %)

The ability to solve the last item depends on successfully solving the first and fourth item. If a student cannot calculate joint probabilities<sup>9</sup>, they will also struggle with applying Bayesian revision. Conversely, if a student can effectively apply Bayesian revision, it indicates proficiency in handling the easier items. This description essentially defines unidimensionality. While the ideal frequency of unidimensionality

violations is expected to be 0 (and  $H = 1$ ), the correlation between items is not expected to be 1.

## Results

The Pearson correlations between Item 5 and the other items are far from reaching 1 (see Figure 4, last column), yet the index  $H$  indicates strong unidimensionality (same Figure, last row). If 2 (out of 133) students were removed from the data set, Item 5 would exhibit perfect unidimensionality with all other items ( $H = 1$ ), but the highest correlation would only reach .52. This outcome underscores that the inequality  $H \geq \phi$  is not merely an esoteric detail but has practical relevance: relying on Pearson correlations can lead to overlooking a unidimensional scale.

**Figure 4:** Pearson correlation and unidimensionality index  $H$  for all item pairs

Item 1	NA	0.89	0.53	0.49	0.27
Item 2	1.00	NA	0.58	0.53	0.31
Item 3	0.81	0.79	NA	0.80	0.41
Item 4	0.86	0.83	0.93	NA	0.49
Item 5	0.86	0.89	0.86	0.88	NA
	1	2	3	4	5

Note. The upper triangular matrix shows the Pearson correlation ( $\phi$ ). The lower triangular matrix shows the unidimensionality index  $H$ .

In line with this statement, interpreting the factor analysis results (Table 4) proves challenging. A one-factor solution only accounts for 40% of the variance in the data, whereas a two-factor solution explains 73%. Clearly, Items 1 and 2 form one factor, Items 3 and 4 another, but Item 5 shows only moderate factor loadings and there are moderate cross-loadings

<sup>8</sup>The actual scenario was in German, and the order of the questions was: 1, 4, 3, 2, 5. Here, the order is shown sorted by the frequency of correct solutions. Note that there were four parallel versions of this test with different numbers in the question text.

<sup>9</sup>Note that the specific method for calculating these probabilities is not crucial. Whether using a formula, drawing a probability tree, or employing another approach, the key point is that without solving Items 1 and 4, Item 5 remains unsolvable.

across all items. To adequately capture the data, a three-factor solution is necessary. Many analysts would likely also explore correlated factors (non-orthogonal rotation). Overall, determining the appropriate approach is not straightforward, even though the data is highly unidimensional. The  $H$  index for all items is .86, clearly surpassing the proposed cutoff of .5 for a strong scale by Mokken (1971, p. 185). However, even this high value likely underestimates the scale's unidimensionality in practical terms. Many violations of unidimensionality can be resolved by examining how participants respond, as demonstrated in the next section.

**Table 4:** Factor analyses for Bayesian revision task

Item	1-factor solution	2-factor solution	
	$\lambda$	$\lambda_1$	$\lambda_2$
1	0.504	0.224	0.915
2	0.583	0.334	0.842
3	0.940	0.903	0.264
4	0.961	0.922	0.281
5	0.494	0.456	0.189

### Explaining Violations of Unidimensionality

In Table 5, the response patterns of the 9 participants who violate unidimensionality are displayed. There are a total of 23 violations, most of which can be categorized and addressed. These violations typically stem from blunders, potential cheating, and conceptual misunderstandings. It is important to emphasize that expertise in the content area is essential for understanding the item response process and resolving violations of unidimensionality.

**Table 5:** Patterns that violate unidimensionality

ID	I1	I2	I3	I4	I5	V	S	explanation
41	0	0	0	0	1	4	1	cheating
108	0	0	1	0	0	2	1	unclear, failed
8	0	0	1	1	0	4	2	red herring
76	0	0	1	1	0	4	2	red herring
89	1	0	1	0	0	1	2	blunder
120	1	0	0	1	0	2	2	unclear, failed
30	0	1	1	1	0	3	3	blunder
36	1	1	0	1	0	1	3	blunder
117	1	1	0	0	1	2	3	cheating

*Note.* ID: Participant ID. I: Item. V: Violations (number of violations). S: Score (the sum score for all five items). The items are sorted by the score.

An interesting category of violations involves blunders, which cannot be entirely avoided but are generally straightfor-

ward to resolve. For instance, Participant 89 used an incorrect probability for Items 2 and 4, entering 91% instead of the correct 81% (based on a base rate of 19%). Despite this error, the participant correctly solved Items 1 and 3, indicating proficiency in calculating joint probabilities. Adjusting for this oversight reduces the number of violations by 1.

Participant 30 provided a value of 12.16% for Item 1, slightly off from the correct value of 12.73%. Given their accurate responses to the subsequent three items, it seems likely that the deviation in Item 1 resulted from minor inattention, such as entering slightly incorrect digits in the calculator. By allowing for a slight tolerance in acceptable solutions, the number of violations is reduced by 3.

Participant 36 also likely made a blunder, answering Item 3 with 27.3% instead of the correct 2.73% (calculated from a base rate of 7% and a conditional probability of 61%:  $0.07(1 - 0.61)$ ). Despite this error, the participant demonstrated the required skill in Items 1, 2 and 4. Correcting this oversight reduces the error count by 1. After addressing potential blunders, 18 violations remain to be explained ( $H = .89$ ).

Another category of violations involves patterns indicative of cheating. For example, Participant 41 failed to solve Items 1 through 4 but surprisingly answered Item 5 correctly. Given the incorrect responses to the earlier items, it seems implausible for the participant to have known the answer to Item 5. Therefore, one approach would be to award no points for Item 5, reducing the violation count by 4.

Similarly, Participant 117 managed to solve only the first two items correctly but not Item 4. It is unclear how the participant could have arrived at the correct solution for Item 5 under these circumstances. Awarding zero points for Item 5 would reduce the violation count by 2. By identifying and addressing these patterns as indicative of cheating, the remaining violation count is reduced to 12 ( $H = .92$ ).

Another group of violations stems from conceptual misunderstandings. Some participants confuse the conditional probability  $P(A|B)$  with the conjunctive probability  $P(A \cap B)$ . This misunderstanding, while somewhat expected, substantially complicates the scaling process. When no specific information is provided, these participants can apply the correct rules. However, when some information is given, they follow a misleading path. Although this occurs very rarely (2 out of 133 participants), it results in a relatively high number of violations (8 in this case).

One potential solution is to modify the task to explicitly clarify the distinction between  $P(A|B)$  and  $P(A \cap B)$ . Another approach could involve explicitly stating that the given probabilities in the text are not the final solution. These modifications might reduce the number of violations. However, there is also a pedagogical argument for retaining the original task, as it highlights areas where students may struggle to differentiate between  $P(A|B)$  and  $P(A \cap B)$ . Addressing these



conceptual misunderstandings is crucial to ensure accurate assessment and understanding among participants.

The two remaining participants (108, 120) exhibit answering patterns that are difficult to explain without further information. Notably, both participants failed the exam, unlike the other seven with violating patterns. Failing an exam often leads to irregular behavior, possibly stemming from motivation issues or time pressure. For instance, participant 120 did not attempt to answer Items 2, 3, and 5, complicating the assessment of these items as “not solved”. One potential approach to gain insights into such behavior could involve filtering out unusual answering patterns post-exam and conducting follow-up interviews with the students. It is crucial that this process does not disadvantage students or affect their grades. The goal is to ascertain whether constructing a perfectly unidimensional scale is possible. With  $H$  already exceeding .9, achieving perfect unidimensionality seems feasible.

Notably, introductions to unidimensionality often lack examples with high  $H$  values. For instance, Ark (2012) present a real data example of a cognitive test for children where a selected subset of seven items achieves  $H = .515$ . Similarly, van Schuur (2011) discusses a religious belief scale with an  $H$  value of .64 as his first real data example. While these values are reasonably high, they miss to fully illustrate the potential of the concept. To convince other researchers of the importance of unidimensionality, it is crucial to showcase examples where  $H$  approaches 1. Such examples clearly demonstrate that achieving perfect unidimensionality is a realistic goal in test development. The Bayesian revision task used in our professorship’s high-stakes exam situation serves as a practical illustration with genuine relevance.

For instance, with a perfectly unidimensional scale, the potential of adaptive testing is significantly enhanced. Solving one specific item implies that easier items would also be solved, thereby optimizing the assessment process. Moreover, detecting cheating patterns could prompt a thorough re-evaluation of exams to identify other irregularities and ensure fairness. Response patterns indicative of specific gaps in knowledge could be systematically analyzed to automatically generate targeted feedback for students, thereby enhancing the quality of teaching and learning outcomes.

This highlights the practical benefits of achieving a perfectly unidimensional scale. However, it is important to note that constructing a unidimensional scale does not guarantee its usefulness, as this primarily depends on the soundness of the underlying theory and various other considerations. Unidimensionality should not replace theoretical work or other criteria in test development, but rather complement them. These points, along with other related ones, are discussed in the final section.

## General Discussion

The aim of this paper was to demonstrate that factor analysis using Pearson correlations ( $\phi$  for dichotomous and polytomous items) is inadequate for assessing dimensional structures. Initially, a simple definition of unidimensionality was explored mathematically, leading to the inequality:  $H \geq \phi$ . This inequality was then applied to construct a theoretical example with dichotomous items, illustrating how factor analysis based on  $\phi$  fails to uncover the true unidimensional structure.

In the empirical section, this issue was investigated using a real-world example featuring five dichotomous items. Despite the items having only moderate correlations, the unidimensionality index  $H$  for the scale was approximately .9. Once again, factor analysis based on  $\phi$  failed to identify a single underlying dimension.

Furthermore, a detailed analysis revealed specific response patterns among the 9 participants who violated unidimensionality. These patterns can potentially indicate blunders, cheating, or specific misunderstandings of the concepts being tested. Psychologists are encouraged not to dismiss this valuable information as noise, but rather to utilize it to enhance the quality of their tests.

While there is clear potential for using  $H$  to improve psychological measurement, psychologists must address numerous other challenges in scaling procedures to achieve satisfactory outcomes. Before delving into these issues, several other aspects of this study warrant discussion.

### When to Use Phi and When to Use H

One potential conclusion from the analysis presented is that perhaps  $\phi$  should simply not be used, at least not in factor analysis. Instead, a viable alternative could be to utilize  $H$  itself, which essentially represents an adjusted version of  $\phi$  (as detailed in the Appendix). However, advocating for the abandonment of  $\phi$  appears short-sighted.

$\phi$  is a widely adopted coefficient due to its exceptional utility. From an informational perspective, it is renowned for its effectiveness, often considered one of the most informative measures for binary classification (Chicco, 2017; Chicco & Jurman, 2023). When predicting responses of participants on an item A based on an item B,  $\phi$  is hard to surpass. It precisely quantifies the success of such predictions: if  $\phi$  equals 1, perfect predictions are achievable.

On the contrary,  $H$  is not suited for this particular purpose. Even if  $H$  equals 1, it does not guarantee accurate predictions. This is because solving an easy item does not necessarily indicate success with a harder item. If most participants solve the easy item, the prediction accuracy is poor, even in a perfectly unidimensional scale. Similarly, the inability to solve a hard item does not imply failure with the easy item.  $H$  primarily signifies unidimensionality—whether items and participants align along a single line. An  $H$  value of 1 simply

indicates this alignment is feasible.<sup>10</sup>

The fundamental question revolves around the researcher's objectives. For instance, if the goal is to identify the most representative item reflecting a latent factor, then factor analysis based on  $\phi$  would be the preferred method, and  $H$  would not be particularly useful. On the other hand, if the aim is to assess the unidimensionality of a scale or identify a subset of unidimensional items,  $H$  appears more suitable for this purpose.

It is important to note that  $H$  can only be computed for dichotomous and polytomous items. For continuous variables, alternative methods are required to assess unidimensionality. For example, implicit association tests utilize reaction times and some cognitive tests also rely on similar measures. Physiological data, being continuous, cannot be analyzed using  $H$ .

While the majority of current psychological tests are dichotomous or polytomous, the inability to calculate  $H$  for continuous data raises the question of whether traditional factor analysis based on Pearson correlations needs to be reconsidered. However, there exists a better alternative that has yet to gain widespread adoption: State-trace analysis (STA, Dunn & Kalish, 2018; Dunn et al., 2019). This approach rigorously tests unidimensionality based on monotonicity and can be applied to continuous variables, albeit currently restricted to experimental factorial data. The theoretical framework of STA is rooted in ordinality, which aligns well with  $H$  and Guttman scaling.

### How to Use H

Ark (2012) has developed an R package that offers a range of functions for calculating  $H$  between items, an item and a scale, and for the entire scale. These functions are particularly valuable when there is a specific hypothesis about which items belong to a single dimension. When conducting factor analysis based on  $\phi$  with a focus on assessing the dimensional structure, it is advisable to verify the results using  $H$ .

The inequality  $H \geq \phi$  suggests that using  $H$  in factor analysis is likely to yield a simpler model, though the improvement may not always be substantial. It is important to recognize that this approach differs from traditional factor analysis using the  $\phi$  coefficient. Factor loadings and scores derived from  $H$  may not be straightforward to interpret, so this method should be used solely to assess unidimensionality.

Ark (2012) also offers a custom implementation utilizing a genetic algorithm to identify unidimensional factors, presenting an alternative to traditional factor analysis. Despite this option, psychologists' familiarity with factor analysis suggests they will be more inclined to integrate  $H$  as an additional input rather than abandoning factor analysis altogether.

A less obvious application of unidimensionality indices is their use in identifying patterns that violate the expected structure. Traditional indices like  $H$  do not directly facilitate

this; instead, such insights come from examining raw Guttman errors. These errors highlight instances where unidimensionality is breached, offering crucial information for test refinement. Guttman (1944, p. 150) summarized this approach succinctly: "In imperfect scales, scale analysis picks out deviants or non-scale types for case studies." Despite its potential, psychologists often neglect analyzing these "deviants", preferring to classify them as mere noise. This tendency hinders a deeper understanding of the underlying item response process. In the context of the Bayesian revision task (refer to Section *Part II: Empirical Example*), the analysis of violations uncovered several ways to improve the test. Yet, there is untapped potential in conducting interviews with participants who defy the scale's expected response patterns.

### The Potential of H for Psychological Science

Researchers who understand the inequality  $H \geq \phi$  can make better-informed decisions, potentially leading to numerous advancements in psychological science. For example, a logical next step would be to re-examine the dimensional structure of existing psychological tests using  $H$ . It is likely that some tests may have fewer dimensions than currently believed, which could simplify the measurement process and the foundational psychological theories on a broad scale.

When developing a new psychological test,  $H$  can be utilized for item selection. Since  $H$  is focused solely on unidimensionality, the reasoning behind including certain items is simplified. For example, very easy and very hard items can be problematic in traditional factor analysis because their variance is restricted, resulting in low factor loadings. However, such items are crucial for distinguishing participants at the extreme ends of the scale. The decision to include or exclude these items often becomes subjective. In such cases,  $H$  could facilitate better decision-making (for a similar point see Sijtsma & Molenaar, 2002, p. 55). If the factor loading based on  $H$  is high, there is an objective reason to include the item, even when the factor loading based on  $\phi$  is low.

Considering the inequality  $H \geq \phi$  during theory development can also be beneficial. Essentially, a perfectly unidimensional scale imposes a systematic restriction on empirical outcomes. A good theory should at least provide a clear item order and explain why certain item responses cannot co-occur. This approach could prove more effective than the currently dominant inductive item selection process, which often relies more on intuition than on logical principles.

Moreover, a sound theory should incorporate the context of item response. The Bayesian revision task, while relatively simple, already reveals complex aspects of this theoretical work. For instance, cheating and guessing are seldom addressed in psychological measurement despite being among

<sup>10</sup>In this scenario, knowing the sum score (total of both items), one can infer responses to both items. However, if only the response to one item is known, such inference is generally not possible.

the most obvious forms of unidimensionality violation. Additionally, participants with very low motivation, such as those anticipating poor performance (e.g. failing an exam), may exhibit erratic response patterns. Currently, these aspects are often dismissed as statistical noise although they can be explained well with plausible psychological processes.

Overall, by re-evaluating dimensional structures and more effectively justifying item selection, researchers can improve the theoretical foundations, ultimately leading to more robust and accurate psychological measurements.

### Alternative coefficients

Although the Pearson correlation is widely used in psychology, alternative measures of association exist that may be more appropriate, especially for dichotomous variables. One such measure is the tetrachoric correlation, which has been recommended by several authors (e.g., Bonett & Price, 2005; Kubinger, 2003; Ledesma et al., 2011). Even Guttman (1944, p. 145) encountered criticism regarding this issue, though he dismissed it by arguing that the tetrachoric correlation assumes continuous and normally distributed latent variables. In contrast, Guttman scaling imposes no such constraints: The latent variable corresponds to the sum score of the scale, which is discrete and does not require adherence to a specific distribution. Indeed, Grønneberg et al. (2020, p. 1040) contended that when the underlying distribution is unknown, “the normal theory tetrachoric correlation coefficient may not be an informative measure of association for binary variables.”

There are additional challenges associated with the tetrachoric correlation, primarily that it cannot be computed directly but must be estimated numerically using various algorithms (e.g. with or without continuity correction). This complexity makes it difficult to interpret and analyze this correlation coefficient. While approximation formulas exist, they may not be entirely satisfactory for rigorous analytical purposes.

Guttman (1944) observed that the tetrachoric correlation does not indicate how well one item can predict another, a similarity it shares with  $H$ . Given this similarity, it would be valuable to explore the relationship between the tetrachoric correlation and  $H$ . However, if they are found to be similar in their applications,  $H$  would be preferable due to its simplicity and ease of interpretation.

For the polychoric correlation, which extends the tetrachoric correlation to polytomous items, similar arguments apply. It introduces additional assumptions that are often difficult to justify, thereby making  $H$  more attractive.

An intriguing alternative to  $H$  is Yule’s  $Q$  for dichotomous variables, equivalent to the  $\gamma$ -coefficient (Goodman & Kruskal, 1954), which can also be applied to polytomous variables. For positive values, similar to  $H \geq \phi$ , it appears that  $Q \geq \phi$  (Eid et al., 2017, p. 556), although the authors do not provide a proof or citation. Preliminary simulations (not shown here)

suggest that the lower bound for  $Q$  exceeds  $\phi$ .  $Q$  shares similarities with an index proposed by Zysno (1993). Based on preliminary simulations, Zysno’s index also exhibits a lower bound larger than  $\phi$ . If Zysno’s index or  $Q$  better reflect unidimensionality compared to  $H$  this could strengthen the arguments presented herein. In particular, the discrepancy between correlation and unidimensionality would be even larger. Future research should delve into the relationships between these coefficients,  $H$  and  $\phi$ .

In the analysis of binary coefficients mentioned earlier (Brusco et al., 2021), Yule’s  $Q$  fell within the same subset as  $\phi$ . Moreover, it is typically classified as a correlation coefficient rather than a measure of unidimensionality. Similarly, the tetrachoric correlation also falls within the same subset as  $\phi$ . Thus, these coefficients still appear distinct from  $H$ , highlighting its unique nature. This distinction is plausible given that  $H$  is conceptualized specifically as a unidimensionality index, not merely a measure of association.

While exploring alternative measures of association and unidimensionality is valuable, the findings presented in this study already bear significant implications. Improving psychological scales can be viewed through the lens of unidimensionality, utilizing  $H$ . Given that unidimensionality is more fundamental than, for instance, reliability or validity, it makes sense to prioritize achieving unidimensionality first and subsequently pursue further refinement. Nevertheless, certain measurement challenges in psychology may pose formidable obstacles to overcome.

### General Measurement Problems

Constructing a good psychological scale typically necessitates a strong theoretical foundation. Simply studying unidimensionality (or factor loadings) and iteratively adjusting the test is unlikely to suffice. In the Bayesian revision task, only five narrowly themed items were included, and it is evident that they should logically form a single dimension. Introducing additional items that appear related could potentially introduce multidimensionality. Success in scale construction often hinges on having a robust theory that guides decisions about which items are suitable or unsuitable for inclusion.

Determining the interdependence of items without a solid theoretical framework is challenging, particularly for tests that do not assess abilities. For instance, in the IPIP-NEO-120 (Johnson, 2014), if a person agrees that they “love big parties”, does this necessarily mean they also agree that they “act wild and crazy”? Conversely, does agreeing with the statement “I act wild and crazy” imply they always “love big parties”? If neither scenario holds true, then these items cannot be regarded as unidimensional measures of extraversion alone; other factors would need to be considered.

Testing abilities is easier because dependencies are more clearly delineated. If an item relies on interim outcomes, the magnitude of  $H$  will be substantial. When a test adheres to

a well-defined dependency structure, a large  $H$  can signify unidimensionality. However, it is possible to design a flawed unidimensional test where interim results are required despite lacking logical dependencies.

For example, solving a maximization problem in mathematics often involves differentiating a function and setting it to zero. While one may know the general concept, the complexity of differentiation, especially when involving rarely used rules, can hinder finding the solution. The resulting ambiguity stems from poorly defined dependencies. Understanding a specific differentiation rule (Item 1) is not a prerequisite for grasping the basic concept of finding maxima of a function (Item 2).

Even with a well-defined dependency tree, another crucial aspect to consider is guessing. If participants can guess the answers,  $H$  cannot be 1. Guessing was intentionally minimized in the Bayesian revision task, but it is often not addressed in many other psychological tests. For instance, most intelligence tests use forced-choice items, which inherently allow for guessing. Reducing guessing in psychological instruments could potentially enhance accuracy without significant additional costs.

Finally, it is important to acknowledge that Guttman scaling, which  $H$  is based on, is ordinal in nature. A scale with  $H = 1$  implies that the sum score reflects perfect order, but it does not quantify distance. This should not be seen as a drawback at the current stage of psychological science. Seeking to establish order in human behavior and experience is a practical objective. In contrast, simply asserting that psychological measurements are interval-scaled does not magically make them interval-scaled (Heene, 2013; Michell, 1990, 2009). A more prudent approach involves constructing strictly unidimensional ordinal scales first, and then aiming for more. Measures like  $H$  are valuable in this pursuit, but the primary challenge lies in refining psychological theories. A sound psychological theory should clearly delineate which empirical outcomes are permissible and which are not. The framework presented in this article has the potential to inspire psychologists to cultivate more refined theories that prominently integrate unidimensionality as a fundamental component, rather than overlooking its importance.

## References

- Ark, L. A. van der. (2012). New developments in mokken scale analysis in R. *Journal of Statistical Software*, 48(5), 1–27. <https://doi.org/10.18637/jss.v048.i05>
- Bonett, D. G., & Price, R. M. (2005). Inferential methods for the tetrachoric correlation coefficient. *Journal of Educational and Behavioral Statistics*, 30(2), 213–225. <https://doi.org/10.3102/10769986030002213>
- Bortz, J., & Schuster, C. (2016). *Statistik für Human- und Sozialwissenschaftler* (7th ed.). Springer.
- Brusco, M., Cradit, J. D., & Steinley, D. (2021). A comparison of 71 binary similarity coefficients: The effect of base rates. *PLOS ONE*, 16(4), e0247751. <https://doi.org/10.1371/journal.pone.0247751>
- Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining*, 10, 35. <https://doi.org/10.1186/s13040-017-0155-3>
- Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, 16(1), 4. <https://doi.org/10.1186/s13040-023-00322-4>
- Choi, S.-S., Cha, S.-H., & Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of systemics, cybernetics and informatics*, 8(1), 43–48.
- Costa, P. T., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment, Vol 2: Personality measurement and testing* (pp. 179–198). Sage Publications. <https://doi.org/10.4135/9781849200479.n9>
- Cureton, E. E. (1959). Note on  $\phi/\phi_{\max}$ . *Psychometrika*, 24(1), 89–91. <https://doi.org/10.1007/BF02289765>
- Davenport, E. C., & El-Sanhurry, N. A. (1991). Phi/Phimax: Review and Synthesis. *Educational and Psychological Measurement*, 51(4), 821–828. <https://doi.org/10.1177/001316449105100403>
- Dunn, J. C., Heathcote, A., & Kalish, M. (2019). Special issue on state-trace analysis. *Journal of Mathematical Psychology*, 90, 1–2. <https://doi.org/10/gn2mjz>
- Dunn, J. C., & Kalish, M. L. (2018). *State-trace analysis*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-73129-2>
- Eid, M., Gollwitzer, M., & Schmitt, M. (2017). *Statistik und Forschungsmethoden* (5th ed.). Beltz.
- Falissard, B. (1999). The unidimensionality of a psychiatric scale: A statistical point of view. *International Journal of Methods in Psychiatric Research*, 8(3), 162–167. <https://doi.org/10.1002/mpr.66>
- Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268), 732–764. <https://doi.org/10.2307/2281536>
- Grønneberg, S., Moss, J., & Foldnes, N. (2020). Partial identification of latent correlations with binary data. *Psychometrika*, 85(4), 1028–1051. <https://doi.org/10.1007/s11336-020-09737-y>
- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9(2), 139–150. <https://doi.org/10.2307/2086306>
- Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behav-*

- ioral Research*, 19(1), 49–78. [https://doi.org/10.1207/s15327906mbr1901\\_3](https://doi.org/10.1207/s15327906mbr1901_3)
- Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9(2), 139–164. <https://doi.org/10.1177/014662168500900204>
- Heene, M. (2013). Additive conjoint measurement and the resistance toward falsifiability in psychology. *Frontiers in Psychology*, 4. <https://doi.org/10.3389/fpsyg.2013.00246>
- Heene, M., Kyngdon, A., & Sckopke, P. (2016). Detecting violations of unidimensionality by order-restricted inference methods. *Frontiers in Applied Mathematics and Statistics*, 2. <https://doi.org/10.3389/fams.2016.00003>
- Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, 51, 78–89. <https://doi.org/10.1016/j.jrp.2014.05.003>
- Kubinger, K. D. (2003). On artificial results due to using factor analysis for dichotomous variables. *Psychology Science*, 45(1), 106–110.
- Ledesma, R. D., Macbeth, G., & Valero-Mora, P. (2011). Software for computing the tetrachoric correlation coefficient. *Revista Latinoamericana de Psicología*, 43(1), 181–189. Retrieved April 17, 2023, from [http://www.scielo.org.co/scielo.php?script=sci\\_abstract&pid=S0120-05342011000100015&lng=en&nrm=iso&tlng=en](http://www.scielo.org.co/scielo.php?script=sci_abstract&pid=S0120-05342011000100015&lng=en&nrm=iso&tlng=en)
- Loevinger, J. (Ed.). (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, 61(4), i–49. <https://doi.org/10.1037/h0093565>
- Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin*, 45(6), 507–529. <https://doi.org/10.1037/h0055827>
- Luce, R. D., & Steingrimsson, R. (2011). Theory and tests of the conjoint commutativity axiom for additive conjoint measurement. *Journal of Mathematical Psychology*, 55, 379–385. <https://doi.org/10.1016/j.jmp.2011.05.004>
- Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1(1), 1–27. <https://doi.org/10/fsf7nv>
- McNemar, Q. (1946). Opinion-attitude methodology. *Psychological Bulletin*, 43(4), 289–374. <https://doi.org/10.1037/h0060985>
- Michell, J. (1990). *An introduction to the logic of psychological measurement*. Psychology Press. <https://doi.org/10.4324/9781315807614>
- Michell, J. (2009). The psychometricians' fallacy: Too clever by half? *British Journal of Mathematical and Statistical Psychology*, 62, 41–55. <https://doi.org/10.1348/000711007X243582>
- Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research*. De Gruyter Mouton. <https://doi.org/10.1515/9783110813203>
- Molenaar, I. W. (1982). Mokken scaling revisited. *Kwantitatieve methoden*, 3(8), 145–164. Retrieved June 28, 2024, from <https://www.vvsor.nl/wp-content/uploads/2020/06/KM1982008013.pdf>
- Molenaar, I. W. (1991). A weighted loevinger H-coefficient extending mokken scaling to multicategory items. *Kwantitatieve Methoden*, 12(37), 97–117. <https://www.vvsor.nl/wp-content/uploads/2020/06/KM1991037007.pdf>
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika*, 74(1), 107–120. <https://doi.org/10.1007/s11336-008-9101-0>
- Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to non-parametric item response theory*. SAGE Publications, Inc. <https://doi.org/10.4135/9781412984676>
- Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research*, 102(3), 443–461. <https://doi.org/10/b8mch3>
- van Schuur, W. H. (2011). *Ordinal item response theory: Mokken scale analysis*. SAGE Publications. <https://doi.org/10.4135/9781452230641>
- van der Ark, L. A., Croon, M. A., & Sijtsma, K. (2008). Mokken scale analysis for dichotomous items using marginal models. *Psychometrika*, 73(2), 183–208. <https://doi.org/10.1007/s11336-007-9034-z>
- Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items. *European Journal of Psychological Assessment*, 31(4), 231–237. <https://doi.org/10/gd52nn>
- Zysno, P. (1993). Polytome Skalogramm-Analyse. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 14(1), 37–49.

## Appendix Derivations

### Derivation 1 (dichotomous)

The first derivation does not rely on any additional theorems and can be performed using only the frequencies  $a, b, c, d$ :

$$H \geq \phi \quad (\text{A1})$$

$$1 - \frac{bN}{(a+b)(b+d)} \geq \frac{ad - cb}{\sqrt{(a+c)(a+b)(b+d)(c+d)}} \quad (\text{A2})$$

Note that all variables are 0 or positive and, as previously described,  $c \geq b$  (the row item is easier or has the same difficulty as the column item). First, we can substitute  $N$  with  $a + b + c + d$  and expand the denominator on the left-hand side:

$$1 - \frac{ba + bc + b^2 + bd}{ab + ad + b^2 + bd} \geq \frac{ad - cb}{\sqrt{(a+c)(a+b)(b+d)(c+d)}} \quad (\text{A3})$$

Now, we can make a single fraction on the left-hand side:

$$\frac{ab + ad + b^2 + bd - ba - bc - b^2 - bd}{ab + ad + b^2 + bd} \geq \frac{ad - cb}{\sqrt{(a+c)(a+b)(b+d)(c+d)}} \quad (\text{A4})$$

And get rid of the same terms:

$$\frac{ad - bc}{ab + ad + b^2 + bd} \geq \frac{ad - cb}{\sqrt{(a+c)(a+b)(b+d)(c+d)}} \quad (\text{A5})$$

The numerators are the same so we can reduce further (note that the inequality holds if  $ad = bc$ ):

$$\sqrt{(a+c)(a+b)(b+d)(c+d)} \geq ab + ad + b^2 + bd \quad (\text{A6})$$

On the right-hand side we can extract the common factors:

$$\sqrt{(a+c)(a+b)(b+d)(c+d)} \geq (b+d)(b+a) \quad (\text{A7})$$

Squaring and reducing further:

$$(a+c)(a+b)(b+d)(c+d) \geq (b+d)^2(b+a)^2 \quad (\text{A8})$$

$$(a+c)(c+d) \geq (b+d)(b+a) \quad (\text{A9})$$

Expanding and reducing further:

$$ac + ad + c^2 + cd \geq b^2 + ba + db + da \quad (\text{A10})$$

$$ac + c^2 + cd \geq b^2 + ba + db \quad (\text{A11})$$

Bringing everything on one side:

$$c^2 - b^2 + ac - ba + cd - db \geq 0 \quad (\text{A12})$$

Factoring out  $c - b$ :

$$(c - b)(c + b) + a(c - b) + d(c - b) \geq 0 \quad (\text{A13})$$

$$(c - b)(c + b + a + d) \geq 0 \quad (\text{A14})$$

$$(c - b)N \geq 0 \quad (\text{A15})$$

Since  $N$  is always positive and  $c \geq b$  (the error frequency is  $b$ ), the inequality  $H \geq \phi$  holds true.

### Derivation 2 (dichotomous and polytomous)

An alternative method of deriving  $H \geq \phi$  is to rewrite  $H$  as a ratio of covariances (see, for instance, Sijtsma & Molenaar, 2002, p. 55):<sup>11</sup>

$$H = \frac{\text{COV}(X, Y)}{\text{COV}(X, Y)^{\max}} = \frac{r(X, Y)}{r(X, Y)^{\max}} \quad (\text{A16})$$

Here,  $X$  and  $Y$  are the two item variables, and  $\max$  indicates the maximum covariance or correlation given the marginal distributions of  $X$  and  $Y$ .<sup>12</sup> Note that the ratio of correlations equals the ratio of covariances because the marginal distributions, and thus the standard deviations of  $X$  and  $Y$ , remain the same for both  $r$  and  $r^{\max}$ . Further note that it suffices to study positive correlations, as a negative correlation between two items can be converted to a positive one by reverse coding one of the items.

Instead of  $r$  one can also use  $\phi$  and rewrite:

$$H = \frac{\phi}{\phi^{\max}} \geq \phi \quad (\text{A17})$$

Because  $\phi^{\max}$  is still a correlation, it cannot exceed 1:

$$\frac{\phi}{\phi^{\max}} \geq \phi^{\max} \quad (\text{A18})$$

$$\phi^{\max} \leq 1 \quad (\text{A19})$$

<sup>11</sup>The following derivation may seem trivial to those familiar with the adjusted  $\phi$  coefficient and its equivalence to  $H$ . When adjusting a positive  $\phi$ , it can only increase, thus it will obviously be at least as large as  $\phi$ . Nonetheless, readers unfamiliar with these concepts can benefit from this formulation.

<sup>12</sup>Specifically,  $r^{\max}$  for the dichotomous case studied here, with a positive correlation and  $p_1 \geq p_2$ , is  $\sqrt{\frac{(a+b)(b+d)}{(a+c)(c+d)}}$ .

By using covariations, this derivation is more general and is also applicable to polytomous items as long as weighted counts of Guttman errors are used (Molenaar, 1991), which is the standard approach (e.g., Ark, 2012; van Schuur, 2011). The meaning of the coefficients remains the same:  $H$  is based on the ratio of Guttman errors to expected errors, while  $\phi$  is simply the Pearson correlation. Although the primary focus of this article is on dichotomous items, it is essential to highlight that the conclusions apply equally to polytomous items.

As a sidenote, it is important to emphasize that  $\frac{\phi}{\phi_{\max}}$  has actually been proposed as an index for dichotomous variables independently of  $H$  (Cureton, 1959). The idea was to create a

version of  $\phi$  that would scale from -1 to 1, regardless of the marginal proportions. Davenport and El-Sanhurry (1991, p. 823) described this index as “a measure of relationship apart from its affiliation with  $\phi$ ”, recognizing that it is not merely an adjustment of  $\phi$  but rather something distinct. Likely unaware of the  $H$  index, they did not make a connection to it. This is also true for several introductions to  $\phi$ , where the max-adjustment is mentioned but not strongly recommended (e.g. Bortz & Schuster, 2016; Eid et al., 2017). While  $\phi$  and the adjusted  $\phi$  are related, their purposes differ, as the adjusted  $\phi$  can be viewed as a measure of unidimensionality.