Journal of Medical Statistics and Informatics
ISSN 2053-7662 | Volume 4 | Article 6

HOAJ
Herbert Open Access Journals

Original Research Open Access

Ranking question designs and analysis methods

Jessica R. Hoag[1,2] and Chia-Ling Kuo[1,2,3*]

*Correspondence: kuo@uchc.edu

CrossMark
Click for updates

[1]Department of Community Medicine and Health Care, University of Connecticut Health Center, USA.
[2]Connecticut Institute for Clinical and Translational Science, USA.
[3]Institute for Systems Genomics, University of Connecticut, USA.

Abstract

Background: Surveys/questionnaires are common research tools to identify the most important barriers that physicians face to improve patient care. Typically, potential items are preselected by investigators and described in a question for participants to rank in order from most to least important. Problems may arise if not every item applies or if multiple items are equally ranked. The nature of this type of ranked data can complicate data analysis and the number of primary items ultimately selected is generally performed ad hoc. To overcome this issue, the question can be broken down into item-specific questions in Likert-style or visual analogue scale (VAS). These two scales are common in psychometric research, but the choice of appropriate statistical methods remains controversial based on the nature and scale of the data.

Methods: In this paper, we compare three question designs (ranking, Likert scale, and VAS) via simulations in order to identify primary items. We focus our investigation on the differences in designs by the scale of data and consider the VAS design as the gold standard as it produces data more informative than the rank and Likert data. We introduce a simulation-based method that accounts for correlation between ranks as well as ratings within subjects.

Results: The VAS design outperforms the Likert and ranking designs. The Likert design with 5 or 7 Likert items is not as compelling as the VAS design until the sample size is large. The ranking design tends to incorrectly identify primary items when the proportion of primary items is over 50%.

Conclusions: Overall, we conclude that the VAS design is the superior choice for identifying and prioritizing primary items. The Likert design can be improved by adding more items. The ranking design is adequate if the proportion of primary items is low.

Keywords: Likert scale, ranking scale, simulation study, questionnaire design, surveys, visual analogue scale, quality improvement

Introduction

Motivation for the present review and investigation originated from the survey question posed to members of the American Orthopaedic Society regarding rates of care for underinsured athletes and barriers influencing the provision of care. Members were asked, "Please rank in order the major barriers for you to provide indigent care (Medicaid, pro bono services) to young athletes in terms of their significance. Rank only those that apply to you". The list of items classified as potential barriers to the provision of care included:

[ ]Lack of time
[ ]Family commitments
[ ]Competing billable hours
[ ]Poor reimbursement
[ ]Costs beyond my time (e.g., OR, hospital, imaging, laboratory)
[ ]Lack of administrative support
[ ]Malpractice insurance concerns
[ ]Employer discouragement
[ ]Very few indigent athletes in my practice community

Results suggested that many participating members only ranked what they considered to be the most significant, or "top" items and assigned the same rank to multiple items. Nonsignificant, or those considered the "least important" items were ranked infrequently and therefore comprised a small sample size due to many missing values. Participants also had unequal contributions to top ranks, with greater contribution from members who gave a top rank to multiple items. For example, one individual may rank all the items as the most important

© 2016 Kuo et al; licensee Herbert Publications Ltd. This is an Open Access article distributed under the terms of Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0). This permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

 doi: 10.7243/2053-7662-4-6

or "top" item and thus have a greater contribution to the top item than the participants who choose to prioritize items. Due to the inconsistent and infrequent ranking of certain items, ranking can be simplified by rating only the top-*u* items where *u* is smaller than the total number of items.

The Friedman's test [1] is a rank-based test alternative to one-way ANOVA with repeated measures. It is equivalent to two-sided Wilcoxon signed rank test when the number of items is two. The Friedman's test can be used to compare the distributions of items and the Skillings-Mack test [2] is a general Friedman's test that handles missing data. An improved test to the Skillings-Mack test recently was published by Best and Rayner [3]. Practically, as suggested by the survey software SurveyMonkey [4]. (http://help.surveymonkey.com/articles/en_US/kb/What-is-the-Rating-Average-and-how-is-it-calculated), researchers tend to summarize the data with mean and standard deviation by item and determine the most important barriers to the provision of care based on the sample means. This method is potentially biased due to what we learned from the data.

Rather than being asked to rank-order items, participants may instead be asked to rate or evaluate each item in a Likert-style scale or visual analogue scale (VAS). The same question can be formatted in Likert or visual analogue scale by asking "Please rate each barrier below for you to provide indigent care (Medicaid, pro bono services) to young athletes in terms of their significance". The two scales differ in the forced rating format. To demonstrate the difference, we use the item, lack of time, as an example, which can be considered in a Likert scale as:

Lack of time:

[ ] Not at all important [ ] Slightly important [ ] Moderately important [ ] Very important [ ] Extremely important

And for visual analogue scale as:

Lack of time:

Not at all important Extremely important

0 _____ 100

Leung [5] classified five types of closed (forced choice) format questions:

1. Choice of categories, e.g., "What is your marital status?" Single, Married, Divorced, Widowed
2. Likert-style scale, e.g., "Statistics is an interesting subject?" Strongly disagree, Disagree, Cannot decide, Agree, Strongly agree
3. Differential scale, e.g., "How would you rate the presentation?" Extremely interesting 1 2 3 4 5 6 7 8 9 10 Extremely dull or a VAS line with extremely interesting and extremely dull at the two ends
4. Checklists, e.g., "Circle the clinical specialties you are particularly interested in?" Accident and emergency, ..., Pediatrics (8 specialties provided)
5. Ranking, e.g., "Please rank your interests in the following specialties?" (1=most interesting, 8=least interesting) Accident and emergency, …, Pediatrics (8 specialties provided)

While our guiding research question stems from an example of barriers to the provision of patient care, the issues encountered with ranking design and its alternatives can be generalized to other applications and research questions informed using survey items. The potential designs in question include 1) ranking through items (ranking the top-*u* items as a special case), 2) rating individual items in a Likert scale, and 3) rating individualitems in a VAS. Leung commented, "*out of these [forced choice] formats, ranking is probably least frequently used, as the responses are relatively difficult to record and analyze*". In fact, the ranking design remains one of the most commonly implemented designs and is assumed in survey software such as SurveyMonkey [4] and SPSS [6].

Questionnaire design is an important topic but is not the focus of this paper. Briefly, Leung [5] and Friedman and Amoo [7] are useful references for a general guideline and a review of VAS can be found in Wewers and Lowe [8]. Typically, a design is evaluated by validity and reliability. In practice, the validity and reliability of individual surveys depend on the research question, study population and study design. For example, the VAS is commonly employed for measurements of pain intensity; the patient is asked to indicate their level of pain on a 10-cm line with rating descriptions of "no pain" on one end and "extreme pain" on the other [9]. The VAS has been shown to be reliable in the acute pain setting, but may exhibit up to 20 percent variability in repeated measurements [9,10]. For Likert-type scales, research on reliability and validity has focused on the ideal number of scale points. In general, valid results can be obtained regardless of the number of scale points, but test-retest reliability is strengthened for scales with between 7 and 10 response categories [11]. In a comparison of ranking and rating scales, Rankin and Grube [12] found no substantial differences in test-retest reliabilities or validity between ranking 2 sets of 18 values and rating the same 36 values on a scale from 1 to 99.

In addition to the difficulties in assessing the most reliable and valid survey measurement method for a given population or research question, there is also debate surrounding the correct methods to analyze the information [13]. Analysis typically involves descriptive statistics to summarize individual items [14]. Ordinal scales are often treated as continuous (i.e., interval) [15,16]. Historically, it has been controversial to presume that the respective differences between ranks or ratings are equal and summarize the data using parametric summary statistics such as mean and standard deviation [17]. In 1946, Stevens [18] argued that parametric statistics should not be applied to ordinal data even if the data is normally distributed because the difference in scales will distort the empirical information [19]. Clason and Dormody [16] similarly suggested that failing to recognize the discrete nature of each item response in a Likert design, while perhaps statistically appropriate, can still lead to inferential errors. They suggested the data be modeled using a multinomial distribution or analyzed using count-based methods, such as chi-square

doi: 10.7243/2053-7662-4-6

tests. Others have discussed the theoretical, empirical, and practical merits of the application of parametric statistics to inherently ordinal scales [19,20]. While parametric statistics come with a larger set of assumptions, namely normality and homoscedasticity, they also provide more power over non-parametric alternatives. Parametric tests for mean comparisons such as the *t*-test and *F*-test are also robust to violations of assumptions [21], a fact that has prompted some to conclude that parametric statistics can be used with Likert data, non-normal distributions, small sample sizes, and unequal variances "with no fear" [13].

In this paper, we compare the VAS, ranking, and Likert designs. Rather than validity and reliability, we focus on the scale of the data produced by the three designs. We simulate VAS scores and convert the data into ranking and Likert scores. The data are analyzed by a simulation-based method that we propose to identify the primary items. We provide an introduction to the simulation-based method and a description of the simulation settings, as well as the simulation results and discussion.

Methods

Notations

In a survey questionnaire, each item is rated by subjects using a VAS, ranking, or Likert score. Assume that k items are under consideration and n subjects participate in the survey. Denote by y_{ij} the VAS score for the j-th item from the i-th subject, $0 \le y_{ij} \le 100$ for $i=1,2,\ldots,n$ and $j=1,2,\ldots,k$. Let the VAS data be collected in the matrix $\boldsymbol{Y}_{(n\times k)}=[\boldsymbol{Y}_1,\boldsymbol{Y}_2,\ldots..,\boldsymbol{Y}_n]^T$ where $\boldsymbol{Y}_i=(y_{i1},y_{i2},\ldots..,y_{ik})^T$ is the vector of VAS score by the i-th subject. Denote by r_{ij} the rank score for the j-th item from the i-the subject, $r_{ij}=k$ if most important, $r_{ij}=1$ if least important. Let the rank score be collected in the matrix $\boldsymbol{R}_{(n\times k)}=[\boldsymbol{R}_1,\boldsymbol{R}_2,\ldots..,\boldsymbol{R}_n]^T$ where $\boldsymbol{R}_i=(r_{i1},r_{i2},\ldots..,r_{ik})^T$ is the vector of rank score by the i-th subject. Denote by s_{ij} the Likert score for the j-th item from the i-th subject. Assume that there are h Likert items, $s_{ij}=h$ if most important, $s_{ij}=1$ if least important. Let the data of Likert score be collected in the matrix $\boldsymbol{S}_{(n\times k)}=[\boldsymbol{S}_1,\boldsymbol{S}_2,\ldots..,\boldsymbol{S}_n]^T$ where $\boldsymbol{S}_i=(s_{i1},s_{i2},\ldots..,s_{ik})^T$ is the vector of Likert score by the i-th subject.

Simulation-based method

The permutation method is a simulation-based method to identify and prioritize top items of the highest importance. This method particularly accounts for correlation between items. For example, in our guiding ranking question: "Please rank in order the major barriers for you to provide indigent care (Medicaid, pro bono services) to young athletes in terms of their significance", the items of family commitments and lack of time are possibly correlated because physicians who have family commitments would tend to have less time to provide indigent care to young athletes. The simulation-based method takes discrete or continuous data, which is described assuming VAS data for convenience. The simulation-based method is an appropriate method for comparison between the three designs because it has the potential to maximally utilize information contained in the data and does not make any distributional assumptions. The procedure is broken down into 8 steps. Steps 1-5 are used to identify top items. Top items are then prioritized to obtain primary items via steps 6-8. To clarify, top items have significantly higher VAS scores than the expected value assuming the items are equally important. top items are a subset of top items and are significantly more important than the other top items. When there is no top item, it implies that the items are equally important. Presumably, the items under investigation are preselected and of high importance. If the score distributions of all items are at the low end, there is no need to perform a test to identify primary items.

Step 1 Let the vector $\boldsymbol{M}^o=(m_{o1},m_{o2},\ldots..,m_{ok})^T$ contain the observed item-specific mean scores where $m_{ok}=\frac{1}{n}\sum_{i=1}^{n} y_{ik}$.

Step 2 Permute the VAS data within subjects, i.e., the data in $\boldsymbol{Y}_i=(y_{i1},y_{i2},\ldots..,y_{ik})^T$ for $i=1,2,\ldots,n$.

Step 3 Collect the data produced at Step 2 and calculate item specific mean scores.

Step 4 Repeat Steps 2 and 3 for B times. Each produces a mean score vector. Denote by $\boldsymbol{M}_i=(m_{i1},m_{i2},\ldots..,m_{ik})^T$ the mean score vector from the i-th permutation where m_{ik} is the mean score of the k-th item at the i-th permutation.

Step 5 The p-value for the j-th item against the null hypothesis that the items are equally important is calculated by

$$p_j = \frac{1}{B}\sum_{i=1}^{B} I(m_{ij} > m_{oj})$$

where $I(m_{ij}>m_{oj})=1$ if $m_{ij}>m_{oj}$; otherwise, $I(m_{ij}>m_{oj})=0$. Items with a p-value smaller than the Bonferroni corrected level, α/k, are considered significant where k is the number of items and α is the overall significance level, typically at 5%. If a single significant item is identified, that item is the only primary item. If more than one significant item is identified and sampling budget is not a concern, these items can all be considered as primary items. Alternatively, significant items can be prioritized using pairwise comparisons conducted sequentially following steps 6-8.

Step 6 Sort the significant items by the observed mean scores, where the top item has the highest observed mean score.

Step 7 Assume that the j'-th and j-th items are the top two significant items and $m_{oj}\ge m_{oj'}$. Denote by $\boldsymbol{D}_{jj'}$ the vector of score difference between the j-th and j'-th items, i.e., $\boldsymbol{D}_{jj'}=\boldsymbol{Y}_j-\boldsymbol{Y}_{j'}$. Sample n differences from the vector with replacement and calculate the mean difference for B times.

Step 8 Let the mean difference from the i-th samples be $\bar{d}_{i;jj'}$. The p-value for the alternative hypothesis that the j-th mean score is greater than the j'-th mean score is calculated by

$$p_{jj'} = \frac{1}{B}\sum_{i=1}^{B} I(\bar{d}_{i;jj'} > 0)$$

where $I(\bar{d}_{i;jj'} > 0) = 1$ if $\bar{d}_{i;jj'} > 0$; otherwise, $I(\bar{d}_{i;jj'} > 0) = 0$

doi: 10.7243/2053-7662-4-6

The *j*-th item is the only primary item if $p_{j\,j'} < \alpha/(k-1)$ where $k-1$ is the maximal number of pair comparisons. Otherwise, compare the top 2 and top 3 items following steps 6-8. The comparison continues until the first significant result is identified or all pairwise comparisons are completed. If the iterative procedure stops at the comparison of the top *j* and *j*+1 items, the top-*j* items are identified as the primary items. If none of the comparisons are significant, the items are considered equally important - all of them are primary items.

Simulations

Since the VAS data contains additional information that is not contained in the rank or Likert data, we first simulate the VAS data and then convert the data to the rank and Likert data. Through the conversion, there is a loss of information in the data. We considered *k=5*, 10 items and *h=5,7* Likert items. Each item was assumed to be either primary or non-primary and the same type of items had a common VAS score distribution. Let the distribution for each primary item be $N(\mu_1,\sigma^2)$ and that for each non-primary item be $N(\mu_0,\sigma^2)$ where $0 \le \mu_0 < \mu_1 \le 100$. VAS scores were simulated jointly for the *k* items from the multivariate normal distribution with the mean vector $[\mu_1\mathbf{1}_{1\times m_1}, \mu_0\mathbf{1}_{1\times m_0}]^T$ and the variance-covariance matrix

$$\sigma^2\begin{bmatrix}\boldsymbol{\Sigma}^1_{m_1\times m_1} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}^0_{m_0\times m_0}\end{bmatrix}$$

where m_1 was the number of primary items and m_0 was the number of non-primary items. $\boldsymbol{\Sigma}^1_{m_1\times m_1}$ and $\boldsymbol{\Sigma}^0_{m_0\times m_0}$ were correlation matrix with 1 on the diagonal, 0 on the non-diagnal for $\boldsymbol{\Sigma}^0_{m_0\times m_0}$, and θ on the non-diagonal for $\boldsymbol{\Sigma}^1_{m_1\times m_1}$. Presumably, primary items tend to be negatively correlated or uncorrelated with non-primary items. Primary items may be correlated with each other; similarly, non-primary items may be correlated with each other. We assumed that any two primary items were correlated with the correlation coefficient θ. Any two non-primary itemswere assumed to have no correlation; similarly, no correlation was assumed between a primary and a non-primary item. Any data point greater than 100 was replaced by 100. Similarly, those smaller than 0 were replaced by 0.

We assumed that the conversion from the VAS data to the rank or Likert data is the same for every participant. We explored multiple conversions with an attempt to minimize information loss. We also handled tied data, which is frequently observed in ranking and Likert data. The idea is similar to simulating data which is extremely non-normal or ideally categorized, i.e., favorable to the ranking and Likert designs. To convert VAS scores to rank scores, we sorted VAS scores within subjects and assigned rank score *k* to the largest and rank score 1 to the smallest. If there was a tie, the average rank score was assigned to each of the items. The conversion was applied to VAS scores before and also after the 0-and-100 truncation. The truncated VAS scores were converted to the same rank scores while VAS scores without being truncated produced unique ranks.

We applied two conversions from VAS scores to Likert scores. Denote by *V* the truncated VAS score and *L* the corresponding Likert score. In the first conversion, we considered the break points $(0, \mu_0-\sigma, \mu_0, \mu_1, \mu_1+\sigma, 100)$ for *h*=5, which formed five intervals:

$$[0, \mu_0-\sigma), [\mu_0-\sigma, \mu_0), [\mu_0, \mu_1), [\mu_1, \mu_1+\sigma), [\mu_1+\sigma, 100]$$

L was assigned *i* if *V* was in the *i*-th interval. Similarly, we considered 6 break points instead for *h*=7,

$$(0, \mu_0-2\sigma, \mu_0-\sigma, \mu_0, \mu_1, \mu_1+\sigma, \mu_1+2\sigma, 100)$$

In the second conversion, we used equal intervals of the length 100/5 for *h*=5 and 100/7 for *h*=7. The first conversion was expected to separate rank scores of primary and non-primary items and give better power than the second conversion.

Simulations with no primary items were termed null simulations. Simulations assuming at least one primary item were termed power simulations. We compared different designs and conversions by type I error rate in null simulations and by power in power simulations. Type I error rate was estimated by the proportion of simulation replicates that incorrectly identified any item when the items were equally important. Power was estimated by the proportion of simulation replicates that correctly identified each item when there was at least one primary item. The simulation parameters are presented in **Table 1**. σ was fixed at 10 for all simulations. In null simulations, $\mu_0 = \mu_1 = \mu$ and we let μ be 80 or 90. All the items are non-primary items, i.e., $m_1 = 0$ and $m_0 = k$. In power simulations, $\mu_0 < \mu_1$. μ_0 was set to 70 or 80 and we chose μ_1 by adding 10 to μ_0. The correlation between primary items (θ) was set to 0, 0.5, 0.8. Number of primary items (m_1) varied from 1 to 4 when *k*=5 and from 1 to 8 when *k*=10. We didn't consider $m_1 = 9$ when *k*=10 because it required a huge sample size to achieve a reasonable power. The results of $m_1 = 1, 2, \cdots, 8$ already showed a trend

Table 1. Simulation parameters.

			Power		Null	
k	*n*	θ	(μ_1, μ_0)	m_1	μ	m_1
5	15	0, 0.5, 0.8	(80, 70), (90, 80)	1	80, 90	0
5	35	0, 0.5, 0.8	(80, 70), (90, 80)	2	80, 90	0
5	80	0, 0.5, 0.8	(80, 70), (90, 80)	3	80, 90	0
5	350	0, 0.5, 0.8	(80, 70), (90, 80)	4	80, 90	0
10	15	0, 0.5, 0.8	(80, 70), (90, 80)	1	80, 90	0
10	25	0, 0.5, 0.8	(80, 70), (90, 80)	2	80, 90	0
10	35	0, 0.5, 0.8	(80, 70), (90, 80)	3	80, 90	0
10	50	0, 0.5, 0.8	(80, 70), (90, 80)	4	80, 90	0
10	80	0, 0.5, 0.8	(80, 70), (90, 80)	5	80, 90	0
10	120	0, 0.5, 0.8	(80, 70), (90, 80)	6	80, 90	0
10	230	0, 0.5, 0.8	(80, 70), (90, 80)	7	80, 90	0
10	500	0, 0.5, 0.8	(80, 70), (90, 80)	8	80, 90	0

doi: 10.7243/2053-7662-4-6

which can be generalized for $m_1 = 9$. The sample size (n) was chosen for a power around 80% assuming VAS data and θ=0. It gave a similar power to (μ_0, μ_1)=(70, 80) and (80, 90). The same sample sizes were used for null simulations regardless of θ. The shift in normal mean would lead to more truncated data. The probability of truncation was 0.1% for data from $N(70, 10^2)$, 2.3% for data from $N(80, 10^2)$, and 15.9% for data from $N(90, 10^2)$. Number of replicates was 1,000 for power simulations (at least one primary item) and was 10,000 for null simulations (items equally important). Number of permutations within each replicate was B=1,000. The overall significance level, α, was set to 5%.

Results

We compare type I error rates and powers of the three designs and conversions from VAS scores to ranks or Likert scores. Specifically, we compare:

1. VAS design (VAS)
2. Rank design with rank scores converted from the original VAS scores (**Rank:O**, Original VAS scores)
3. Rank design with rank scores converted from the truncated VAS scores (**Rank:T**, Truncated VAS scores)
4. Likert design with 5 Likert items where Likert scores are obtained via the first conversion based on the intervals formed by population means and standard deviations of primary and non-primary items (**L5:UEI**, Unequal Intervals)
5. Likert design with 7 Likert items where Likert scores are obtained via the first conversion based on the intervals formed by population means and standard deviations of primary and non-primary items (**L7:UEI**, Unequal Intervals)
6. Likert design with 5 Likert items where Likert scores are obtained via the second conversion based on 5 equal intervals (**L5:EI**, Equal Intervals)
7. Likert design with 7 Likert items where Likert scores are obtained via the second conversion based on 7 equal intervals. (**L7:EI**, Equal Intervals)

Null simulations

In null simulations, $\mu_1 = \mu_0 = \mu$, i.e., no primary items, where $\mu = 80, 90$. The null simulation results are presented in **Figure 1** for k=5 (5 items) and in **Figure 2** for k=10 (10 items). The y-axis denotes type I error rate. The sample sizes on the x-axis are the sample sizes chosen for power simulations to achieve a power around 80% assuming a number of uncorrelated primary items. The left three panels are for μ=80 with the top one for θ=0 followed by θ=0.5 and then θ=0.8. Similarly, the right three panels are for μ=90 and θ=0, 0.5, 0.8.

Figures 1 and **2** show consistent results. In each figure, the patterns are similar regardless of the correlation between primary items (θ). With the exception of L5:EI and L7:EI, simulation designs are robust to the data truncation and perform similarly for μ=80, 90. The type I error rates of L5:EI and L7:EI are inflated by the truncation and the inflation is intensied when the sample size is small and, for L5:EI only, when the number of items is 10. The VAS consistently results in a reasonable type I error in the 95% confidence interval of 5% nominal level, [0.0457, 0.0543]. Other designs have an inflated type I error rate when the sample size is small. The inflation diminishes when the sample size is sufficiently large. The degree of inflation and the sample size needed to maintain a reasonable type I error, however, varies with designs and conversions. The most inflated type I error rate occurs when the sample size is 15 (the smallest sample size that we considered in our simulations) and when the data is greatly truncated for L5:EI or L7:EI. The type I error rates are 0.0555 for VAS, 0.0679 for RANK:O, 0.0655 for RANK:T, 0.0721 for L5:UEI, 0.0701 for L7:UEI, 0.5424 for L5:EI, and 0.1473 for L7:EI. In **Figures 1** and **2**, the type I error rates of L5:EI when the sample size is 15 are relatively large and are omitted in order to appropriately scale the figures. Overall, VAS design performed the best followed in descending order by RANK:O, RANK:T, L5:UEI, L7:UEI, L7:EI, and L5:EI. Within designs, L7:EI consistently performs better than L5:EI. RANK:O and RANK:T behave similarly, as do L5:UEI and L7:UEI.

Power simulations

The power simulation results are presented in **Figures 3** and **4**, **Figure 3** for k=5 (5 items) and **Figure 4** for k=10 (10 items). Both figures are arranged by the correlation between primary items and the population means of primary and non-primary items. The x-axis denotes the number of primary items and the y-axis denotes power.

From k=5 to k=10, the patterns of power are similar. When the number of primary items is increased, the power of every design and conversion increases with the exception of RANK:O and RANK:T. Both continue to maintain a power approximately equal to 0.80, but start losing power when the proportion of primary items is greater than 50%. With the increase in correlation between primary items (θ), the power of every design and conversion increases, but this increase is not as significant for RANK:T and RANK:O. Other than L5:EI and L7:EI, simulation designs are robust to the right shift of normal means where 15.9% of VAS scores for each primary item are truncated at 100. The power of L5:EI and L7:EI, respectively, is reduced, particularly L5:EI, when k=10 and the proportion of primary items is small. The VAS design consistently demonstrates the highest power, followed by L5:UEI, L7:UEI, L7:EI, L5:EI, RANK:O, and RANK:T. L5:UEI and L7:UEI show consistently similar power, while L7:EI is consistently more powerful than L5:EI. RANK:O and RANK:T perform similarly with better power than L5:EI when the proportion of primary items is smaller than 50%.

Discussions

We have explored three survey questionnaire designs (ranking, Likert, and VAS) for their ability to identify primary items. The ranking design asks participants to weigh different items simultaneously, whereas the Likert and VAS designs require participants to evaluate items one at a time. While making a fair

doi: 10.7243/2053-7662-4-6

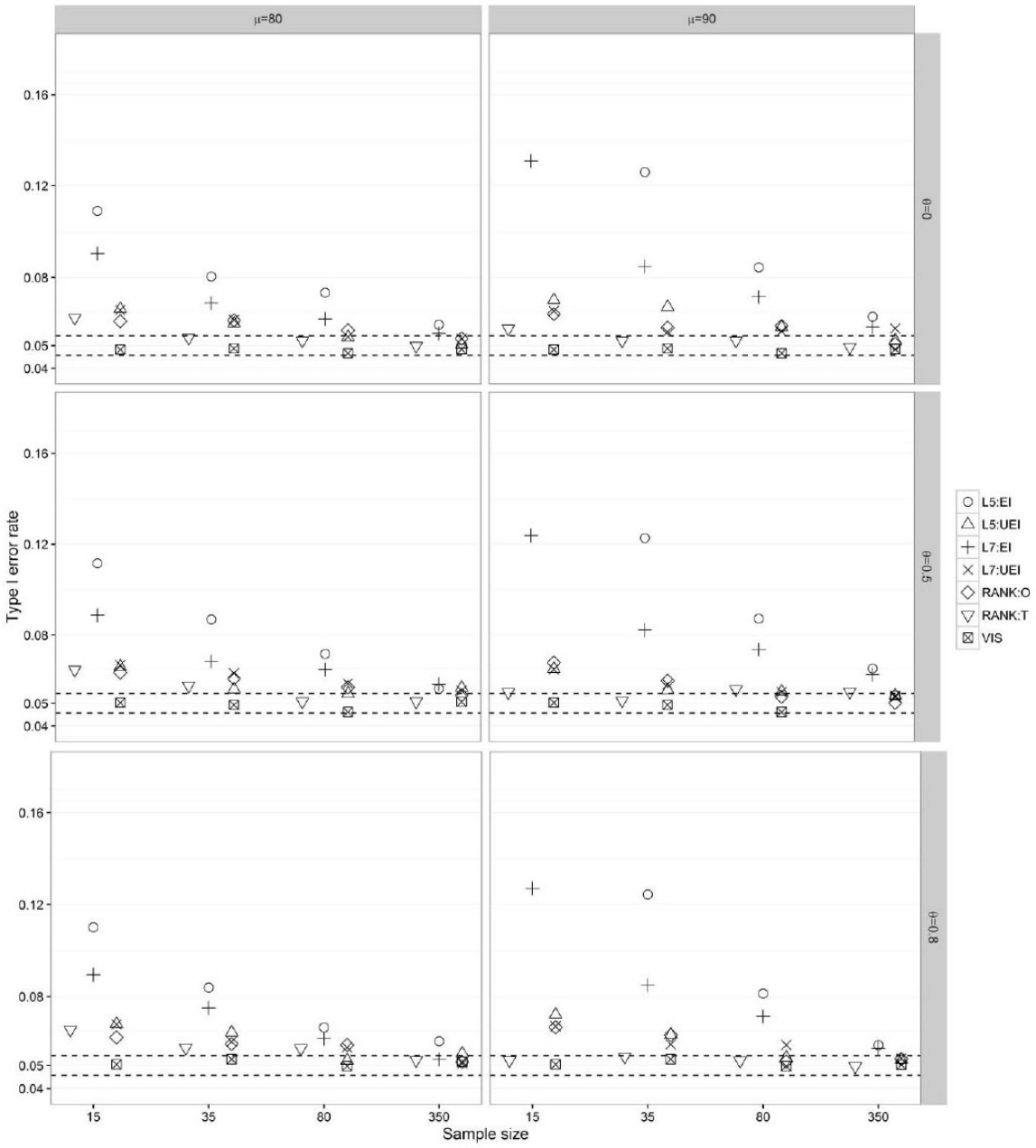

Figure 1. Null simulation results: type I error rates assuming the number of items is 5 (the two dash lines are y=0.0457 and y=0.0543, the lower and upper bounds of 95% confidence interval of 5% nominal level). VAS-VAS design, Rank:O - rank design with rank scores converted from the original VAS scores, Rank:T - rank design with rank scores converted from the truncated VAS scores, L5:UEI - Likert design with 5 Likert items where Likert scores are obtained via the conversion based on the intervals formed by population means and standard deviations of primary and non-primary items, L7:UEI - Likert design with 7 Likert items where Likert scores are obtained via the conversion based on the intervals formed by population means and standard deviations of primary and non-primaryitems, L5:EI - Likert design with 5 Likert items where Likert scores are obtained via the conversion based on 5 equal intervals, L7:EI - Likert design with 7 Likert items where Likert scores are obtained via the conversion based on 7 equal intervals.

comparison of the three design scales is a challenge, we were able to introduce a simulation-based method and evaluation scheme to determine the probability of correctly identifying each item, primary or non-primary. Presumably, the preselected items are of high importance. Therefore, we simulated VAS scores at the high end and studied the effect of truncation

doi: 10.7243/2053-7662-4-6

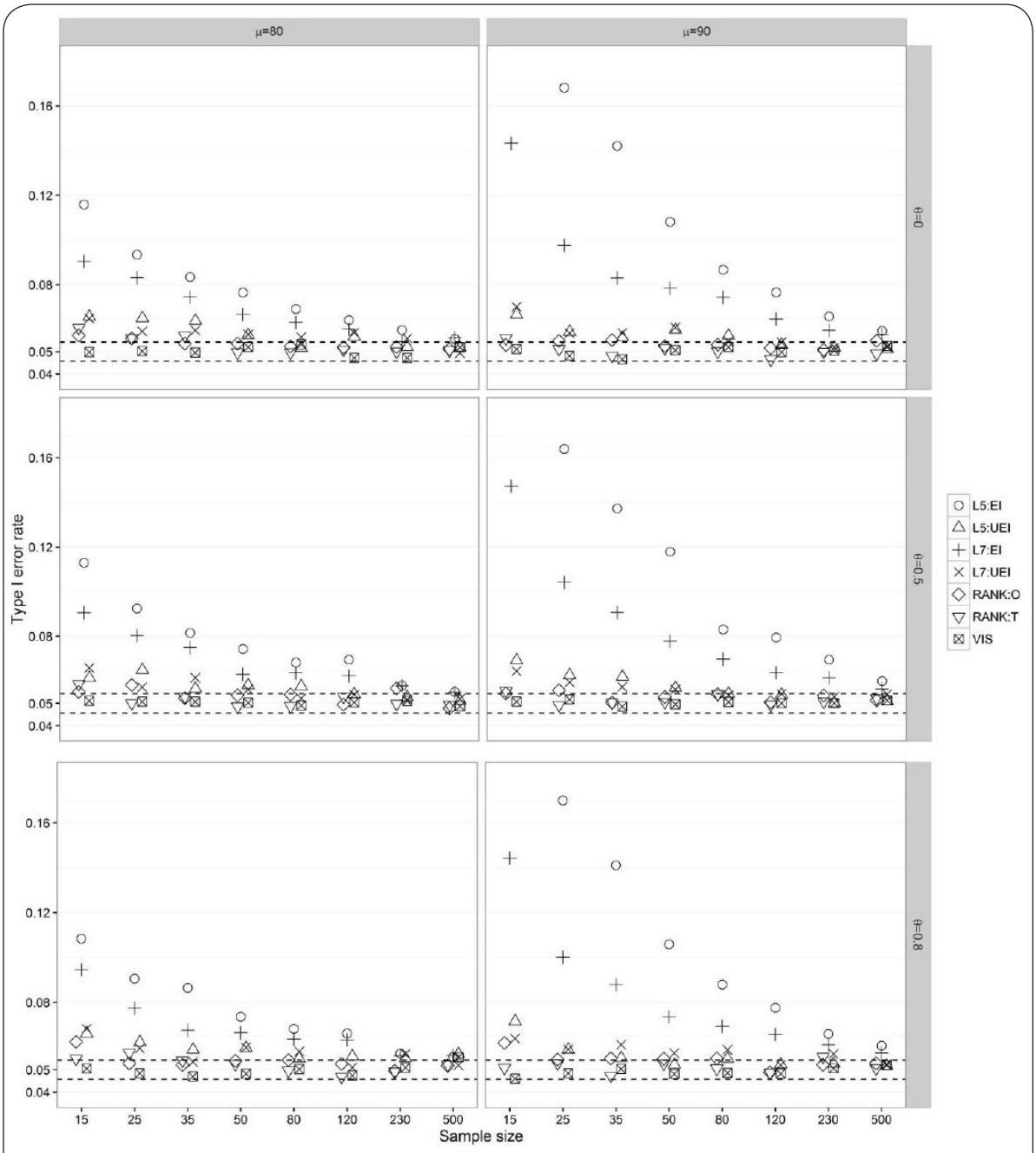

Figure 2. Null simulation results: type I error rates assuming the number of items is 10 (the two dash lines are y=0.0457 and y=0.0543, the lower and upper bounds of 95% confidence interval of 5% nominal level). VAS-VAS design, Rank:O - rank design with rank scores converted from the original VAS scores, Rank:T - rank design with rank scores converted from the truncated VAS scores, L5:UEI - Likert design with 5 Likert items where Likert scores are obtained via the conversion based on the intervals formed by population means and standard deviations of primary and non-primary items, L7:UEI - Likert design with 7 Likert items where Likert scores are obtained via the conversion based on the intervals formed by population means and standard deviations of primary and non-primary items, L5:EI - Likert design with 5 Likert items where Likert scores are obtained via the conversion based on 5 equal intervals, L7:EI - Likert design with 7 Likert items where Likert scores are obtained via the conversion based on 7 equal intervals.

or a left-skewed distribution. To ensure comparable data across designs, we began with VAS scores and converted them to ranking and Likert scores. Multiple conversions were considered for each design, each having a different practical implication.

doi: 10.7243/2053-7662-4-6

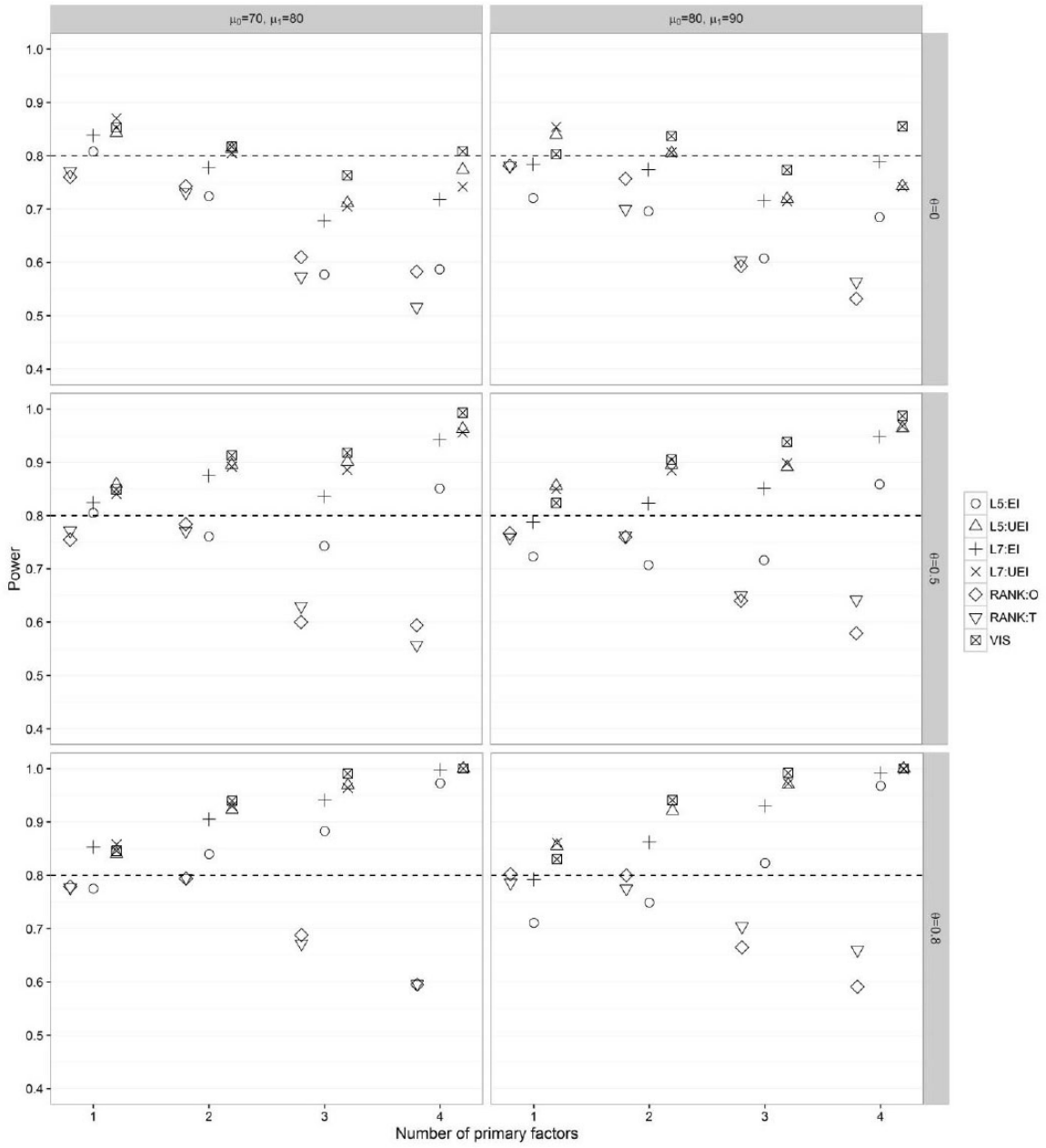

Figure 3. Power simulation results: powers assuming the number of items is 5. VAS-VAS design, Rank:O - rank design with rank scores converted from the original VAS scores, Rank:T - rank design with rank scores converted from the truncated VAS scores, L5:UEI - Likert design with 5 Likert items where Likert scores are obtained via the conversion based on the intervals formed by population means and standard deviations of primary and non-primary items, L7:UEI - Likert design with 7 Likert items where Likert scores are obtained via the conversion based on the intervals formed by population means and standard deviations of primary and non-primary items, L5:EI - Likert design with 5 Likert items where Likert scores are obtained via the conversion based on 5 equal intervals, L7:EI - Likert design with 7 Likert items where Likert scores are obtained via the conversion based on 7 equal intervals.

From VAS to ranking scores, the conversion to untruncated VAS scores assumes that each item is uniquely ranked unless VAS scores are exactly the same -- the likelihood of this similarity is nearly 0. The conversion to truncated VAS scores allows the top items to be equally ranked. This conversion can be generalized to items with similar VAS scores being assigned the same rank scores. In both conversions, the rank scores for items with equal ranks are obtained by taking the average of rank scores presuming they are uniquely ranked and the sum of rank scores for all items is the same for each subject, i.e., each

doi: 10.7243/2053-7662-4-6

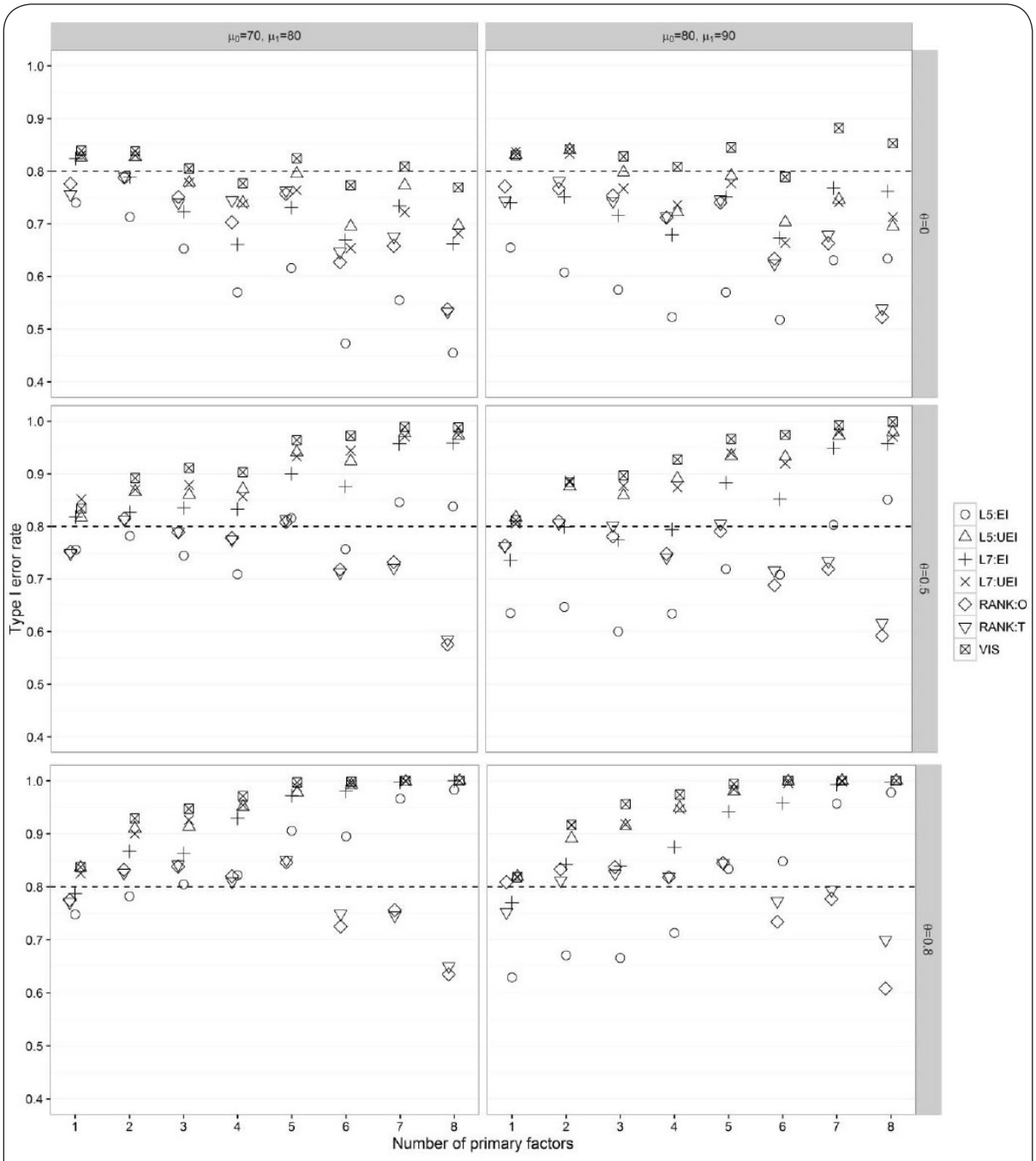

Figure 4. Power simulation results: powers assuming the number of items is 10. VAS - VAS design, Rank:O - rank design with rank scores converted from the original VAS scores, Rank:T - rank design with rank scores converted from the truncated VAS scores, L5:UEI - Likert design with 5 Likert items where Likert scores are obtained via the conversion based on the intervals formed by population means and standard deviations of primary and non-primary items, L7:UEI - Likert design with 7 Likert items where Likert scores are obtained via the conversion based on the intervals formed by population means and standard deviations of primary and non-primary items, L5:EI - Likert design with 5 Likert items where Likert scores are obtained via the conversion based on 5 equal intervals, L7:EI - Likert design with 7 Likert items where Likert scores are obtained via the conversion based on 7 equal intervals.

subject has the same contribution to the data. Whether there are ties or not, the mean scores of neighboring items will be similar and the comparisons between items will not be affected. From VAS to Likert scores, the conversion based on equal intervals is quite straightforward but the resulting distributions of Likert scores for primary and non-primary items may

not be apart, particularly when the number of Likert items is 5. The conversion based on unequal intervals assumes a level of item importance. These intervals are data dependent and are determined such that the Likert scores of primary and non-primary items are well separated. If real data are available, the relationship between VAS and Likert scores is expected to be reasonably described by the first conversion. The second conversion produces data by attempting to optimize the Likert design. It implies a design with numerous Likert items where the bottom items are truncated and the differences between items are unequal.

Our simulation results show that the VAS design consistently has a reasonable type I error rate, as well as the highest power when compared to ranking and Likert designs. Unsurprisingly, L5:UEI and L7:UEI outperform L5:EI and L7:EI. L5:UEI and L7:UEI are as competitive as VAS except when the primary items are uncorrelated and the proportion of primary items is high. L5:UEI and L7:UEI perform similarly, which implies that prioritizing the extreme values does not necessarily improve the performance of Likert design. L7:EI consistently outperforms L5:EI. Both have a highly inflated type I error when the sample size is not sufficiently large. A large sample size is required to correctly identify each item when the proportion of primary items is high. When the validity (type I error rate) is not a problem and the primary items are correlated, L7:EI and VAS are similarly powerful. UEI designs utilize the distributional information and are favored over EI designs with the exception of L7:UEI, which has a higher power than L7:EI when items are independent, both designs are empowered (with a sufficiently large sample size), and the data is substantially right truncated. With the increase in sample size, EI designs are able to maintain a reasonable type I error and improve power. UEI becomes a less sensible method to convert VAS to Likert data compared to EI when the data is substantially right truncated and the number of Likert items is 7.

For the Likert design, we have seen significant improvement in performance from 5 to 7 Likert items. Still, 7 Likert items are insufficient for the Likert score distribution to serve as a good approximation to the VAS score distribution. More Likert items are needed to improve the Likert design. Rather than describe different degrees of importance in words, it is a better idea to group VAS scores in 10 groups, for example, and ask participants to check one of the groups, 1-10, 11-20, …, and 91-100. It is essentially adding more items to improve the approximation to a VAS score distribution.

The superiority of the VAS design over the ranking design is minimized when there are very few primary items. The ranking design may allow for ties but that is not related to the identification of primary items. This design is robust to data truncation and correlation between primary items. It has a slightly inflated type I error rate when the sample size is not sufficiently large. The power significantly drops when the number of primary items is greater than the number of non-primary items. The condition that improves EI designs reduces the power of ranking design - an increase in the number of primary items. Assume four out of five items are primary items, items A, B, C, and D. A is consistently ranked the top 1, B the top 2, C the top 3, and D the top 4. The average rank score is 3 and the item D will never be identified a primary item. Similarly, if there are 10 items with the assumption that each primary item is ranked the same by participants. The average rank score is 5.5 and primary items ranked 6 or greater will never be identified. With the increase in correlation between primary items, the scores of primary items become consistent, i.e., the likelihood that any primary item has a low score like one from the distribution of a non-primary item is reduced. This results in an increase in power. The ranking design does not benefit from this increased correlation because regardless of how correlated two items are, it produces ties and the power will not change (explained previously).

While the simulation-based method is introduced to compare different designs, it can also be used to decide which items to carry forward for intervention. This method is a two-stage approach. The first stage identifies top items and the second stage prioritizes them. Both stages provide information for decision-making. The simulation-based method is simple and accounts for the correlation between scores within subjects. It is particularly useful when the distributions overlap. The simulation-based method performs best for VAS data. The ranking and Likert designs are outperformed due to the scale of data. While the debate regarding analysis methods for survey data continues, most of the empirical literature suggests that the choice between parametric vs. non-parametric methods will not bias the validity of results [**13**,**22**]. While we simulated truncated data, we did not explore extremely non-normal data. Instead, we explored multiple conversions to minimize information loss and tied data, which is frequently observed in ranking and Likert data. The concept is similar to simulating data which is extremely non-normal or ideally categorized, i.e., favorable to the ranking and Likert designs. Extremely non-normal data may fail the simulation-based method by showing no preference among the designs. It will never show that the ranking or Likert design is better than the VAS design because the VAS data is more informative than the rank and Likert data. Ideally, a better method that is able to utilize most of the data information, e.g., Friedman's test, then should be used to compare the three designs. We utilized the Bonferroni method for multiple testing correction in the simulation-based method. The Bonferroni method has been cited as conservative and other methods such as Holm's Bonferroni procedure [**23**] to control family wise error rate in a less conservative way may be used instead to improve the power. In addition, the simulation-based method can be adapted to control false discovery rate instead of family wise error rate. Methods to control false discovery rate include those proposed by Benjamini and Hochberg [**24**]. For general use, how this simulation-based method can be adapted for missing data and how robust it is to outliers and extremely

doi: 10.7243/2053-7662-4-6

non-normal distributions, however, requires future investigation. The Friedman's type tests can be an alternative to the simulation-based test for the design comparison, particularly the Skillings-Mack test is readily applied to study missing data. These tests are robust to non-normal data; however, are conservative against the VAS design when the data is approximately normal.

Conclusions

Overall, we conclude that the VAS design is the superior choice for identifying and prioritizing primary items. A Likert design, however, will perform just as well when the sample size is large. The Likert design can also be improved by adding more items. The ranking design tends to incorrectly identify primary items, but it is adequate if the proportion of primary items is low. In addition to the scale of data, other considerations need to be taken into account before choosing a design. For example, certain populations (i.e., children or intellectually disabled) find the VAS more difficult to understand and answer correctly [**25**]. Additionally, as surveys are more often delivered online, the VAS design may not be an option due to implementation difficulties such as not being user-friendly or creating many missing values [**26**]. As we know, SurveyMonkey [**4**] currently does not include a function for users to create VAS questions online in the way that one can indicate a position by mouse click along a continuous line between two end-points. Thus, several variants of the VAS design have been practiced including allowing a user to enter a score [0-100], instead of marking a point on a measurement line. Although we have not considered these variants in our study, they are likely to share similarities with VAS and Likert designs from an analysis prospective. There is some evidence that software companies are overcoming the operational difficulties associated with VAS implementation in the form of slider bars [**27**], and it is our hope that web survey companies continue to expand the extent to which the VAS design is available for users to implement their survey questions.

List of Abbreviation

VAS: Visual analogue scale

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Authors' contributions	JRH	CLK
Research concept and design	✓	✓
Collection and/or assembly of data	--	--
Data analysis and interpretation	✓	✓
Writing the article	✓	✓
Critical revision of the article	✓	✓
Final approval of article	✓	✓
Statistical analysis	--	✓

Acknowledgement and funding

We thank the reviewers' suggestions and comment that greatly improved the quality of this paper.

Publication history

Editor: Nicola Shaw, Algoma University, Canada.
EIC: Jimmy Efird, East Carolina University, USA.
Received: 14-Jun-2016 Final Revised: 12-Jul-2016
Accepted: 29-Jul-2016 Published: 10-Aug-2016

References

1. M. Friedman. **The use of ranks to avoid the assumption of normality implicit in the analysis ofvariance**. *Journal of the american statistical association*. 1937; **32**:675-701. | Pdf
2. J.H. Skillings and G.A. Mack. **On the use of a Friedman-type statistic in balanced and unbalanced block designs**. *Technometrics*. 1981; **23**:171-177.
3. D. J. Best and J. C. W. Rayner. **Analysis of ranked data in randomized blocks when there are missing values**. *Journal of Applied Statistics*. 2016; 1-8. | Article
4. A. Gordon. **Surveymonkey**. *The Internet and Higher Education*. 2002; **5**:83-87. | Website
5. W.-C. Leung. **How to design a questionnaire, student**. *BMJ*. 2001; **9**:187-189.
6. IBM. **IBM SPSS Data Collection Survey Reporter 6.0.1 Users Guide**. 2011; **1**.
7. H. H. Friedman and T. Amoo. **Rating the rating scales**. *Journal of Marketing Management*. 1999; 114-123.
8. Wewers ME and Lowe NK. **A critical review of visual analogue scales in the measurement of clinical phenomena**. *Res Nurs Health*. 1990; **13**:227-36. | Article | PubMed
9. Williamson A and Hoggart B. **Pain: a review of three commonly used pain rating scales**. *J Clin Nurs*. 2005; **14**:798-804. | Article | PubMed
10. Carlsson AM. **Assessment of chronic pain. I. Aspects of the reliability and validity of the visual analogue scale**. *Pain*. 1983; **16**:87-101. | PubMed
11. Preston CC and Colman AM. **Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences**. *Acta Psychol (Amst)*. 2000; **104**:1-15. | Article | PubMed
12. W. L. Rankin and J. W. Grube. **A comparison of ranking and rating procedures for value system measurement**, *European Journal of Social Psychology*. 1980; **10**:233-246. | Article
13. Norman G. **Likert scales, levels of measurement and the "laws" of statistics**. *Adv Health Sci Educ Theory Pract*. 2010; **15**:625-32. | Article | PubMed
14. Carifio J and Perla R. **Resolving the 50-year debate around using and misusing Likert scales**. *Med Educ*. 2008; **42**:1150-2. | Article | PubMed
15. Kersten P, Kucukdeveci AA and Tennant A. **The use of the Visual Analogue Scale (VAS) in rehabilitation outcomes**. *J Rehabil Med*. 2012; **44**:609-10. | Article | PubMed
16. D. L. Clason and T. J. Dormody. **Analyzing data measured by individual likert-type items**. *Journal of Agricultural Education*. 1994; **35**:4. | Pdf
17. S. Jamieson. **Likert scales: how to (ab) use them**. *Medical education*. 2004; **38**:1217-1218. | Pdf
18. Stevens SS. **On the Theory of Scales of Measurement**. *Science*. 1946; **103**:677-80. | Article | PubMed
19. Armstrong GD. **Parametric statistics and ordinal data: a pervasive misconception**. *Nurs Res*. 1981; **30**:60-2. | Article | PubMed
20. Knapp TR. **Treating ordinal scales as interval scales: an attempt to resolve the controversy**. *Nurs Res*. 1990; **39**:121-3. | Article | PubMed
21. L. L. Havlicek and N. L. Peterson. **Robustness of the t test: A guide for researchers on effect of violations of assumptions**. *Psychological Reports*. 1974; **34**:1095-1114. | Article
22. Maxwell C. **Sensitivity and accuracy of the visual analogue scale: a psycho-physical classroom experiment**. *Br J Clin Pharmacol*. 1978; **6**:15-24. | Article | PubMed Abstract | PubMed FullText

23. S. Holm. **A simple sequentially rejective multiple test procedure**. *Scandinavian journal of statistics*. 1979; 65-70. | Pdf

24. Y. Benjamini and Y. Hochberg. **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society Series B*. 1995; **57**:289-300. | Article

25. D. Hasson and B. B. Arnetz. **Validation and findings comparing vas vs. likert scales for psychosocial measurements**. *International Electronic Journal of Health Education*. 2005; **8**:178-192. | Pdf

26. M. P. Couper, R. Tourangeau, F. G. Conrad and E. Singer. **Evaluating the effectiveness of visual analog scales a web experiment**. *Social Science Computer Review*. 2006; **24**:227-245. | Article

27. Survey Analytics. **Standard Question Types**. 2016. | Website

Citation:
Hoag JR and Kuo C-L. **Ranking question designs and analysis methods**. *J Med Stat Inform*. 2016; **4**:6.
http://dx.doi.org/10.7243/2053-7662-4-6