INTERNATIONAL DESIGN CONFERENCE - DESIGN 2016
Dubrovnik - Croatia, May 16 - 19, 2016.

TOWARD A COMMON PROCEDURE USING LIKERT AND LIKERT-TYPE SCALES IN SMALL GROUPS COMPARATIVE DESIGN OBSERVATIONS

A. L. Guerra, T. Gidel and E. Vezzetti

Keywords: computer aided design (CAD), design activities, design research, research methodologies and methods, experiment

1. Introduction

Creating, introducing, and assessing new computer supports is a common research subject in the design research community [Blessing 1994], [Maher et al. 1998], [Hinds and Kiesler 2002], [Tang et al. 2011]. A widespread procedure is based on comparative methods: "*the selection and analysis of cases that are similar in known ways and differ in other ways, with a view to formulating or testing hypothesis*" [Jupp 2006, p.33]. Buisine et al. label this procedure as paradigm evaluation [Buisine et al. 2012], which consists in comparing independent groups (very often small groups [Bale 1950]) realizing the same activity on a computer support and on a given control condition (e.g. pen-and-paper). Different factors are measured in paradigm evaluation (e.g. concept generation or creativity [Gidel et al. 2011], distal or located work [Tang et al. 2011], collaboration [Buisine et al. 2012]). However, a common investigation concerns users' opinions toward the perceived quality of new computer supports (e.g perceived usability, perceived effectiveness, perceived efficiency) [Blessing 1994], [Guerra et al. 2015].

Frequently, users' opinions are measured through questionnaires based on Likert and Likert-type items, forming rating scales. The resulting data are analyzed in order to find any difference whether positive or negative between the two conditions. Hence, a common but incorrect practice is to categorize results as "good" or "bad" according to their statistical significance (i.e. p-value). The statistical analysis of this data, and so, the way statistical significance is calculated, depends on the rating scale type employed. However, confusion persists about the correct classification of rating scales, whether they are Likert, Likert-type scales or Discrete Visual Analog Scales (DVAS). This may seems a trivial problem, but this incertitude hinders the comparison of existing research, causing a lack of scientific rigor and lowering the impact of the results for practical use. Section 2 clarifies this terminological debate with a deductive approach: from Likert scales to general rating scales.

How to use Likert and Likert-type scales is still an open question [Jamieson 2004], [Norman 2010], in particular do we need to statistically threat them as ordinal or interval scales? Section 3 will presents the different points of view on the matter with a numerical example.

Section 4 propose a pragmatic solution and goes further, delving deeper into the issue if is correct to use or not p-value as a divide between "good" and "bad" results. Finally, section 5 proposes guidelines to be applied when using Likert and Likert-type scales in comparative design observations of small groups. This proposition, open to be debated, aim to provide a pragmatic common procedure that increase scientific rigor and comparability of design research.

2. From rating scales to Likert scales

The first aspect that needs to be clarified is terminology. Over time and in common usage, the term, "Likert scale" has come to be applied to things far removed from its original meaning, such as Likert-type scales, discrete visual analog scale, and rating scales [Uebersax 2006].
Figure 1 proposes a framework describing the hierarchical relations among these different scales.

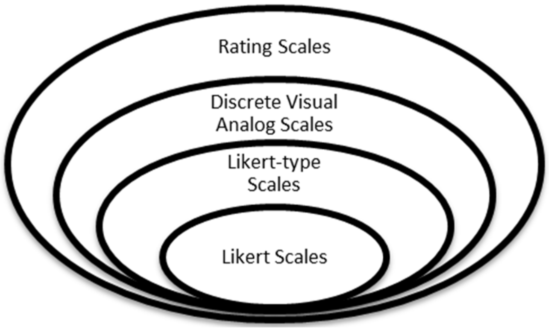


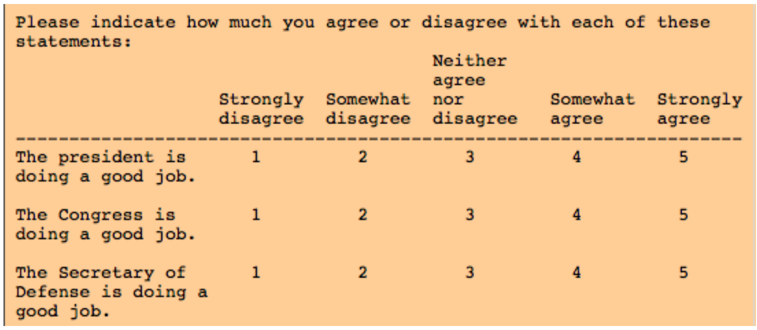
Figure 1. Hierarchical relation between Rating scales, Discrete Visual Analog scales, Likert-type scales, and Likert scales

2.1 Likert scales

Attitude scales (later named Likert scales) have been introduced by Rensis Likert in June 1932, with the work "A technique for the measurement of attitudes" [Likert 1932]. Moreover these attitudes are conceived as clustered, so that a group factor is assumed as outset [Likert 1932, p.10]. This is the reason for which Likert-scales are defined as multi-Likert items scales.
A Likert item is a single statement composing a Likert-scale. It should respect the following characteristics [Uebersax 2006]:

1. Response levels are anchored with consecutive integers;
2. Response levels are arranged horizontally.
3. Response levels are also anchored with verbal labels which connote more-or-less evenly-spaced gradations.
4. Verbal labels are bivalent and symmetrical about a neutral middle (consequently there is usually an odd number of response levels, 5,7,9,11).
5. In Likert's usage, the scale always measures attitude in terms of level of agreement/disagreement to a target statement.

An example of a Likert scales composed by 3 Likert items is presented in Figure 1



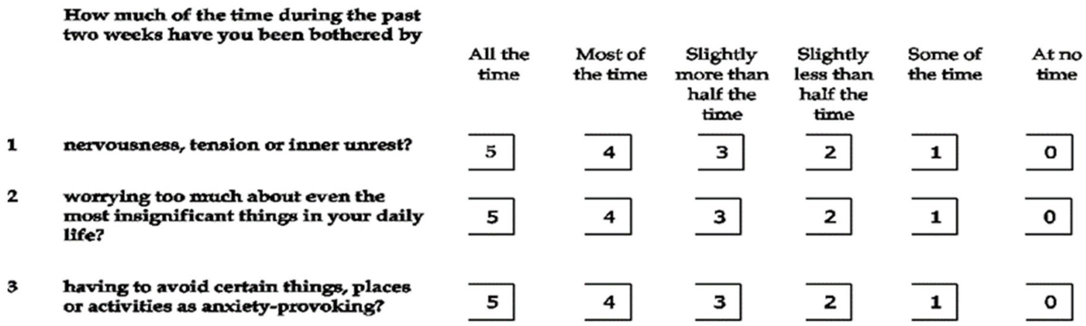
Please indicate how much you agree or disagree with each of these statements:

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
The president is doing a good job.	1	2	3	4	5
The Congress is doing a good job.	1	2	3	4	5
The Secretary of Defense is doing a good job.	1	2	3	4	5

Figure 2. Example of Likert scale (from http://john-uebersax.com/stat/likert.htm)

2.2 Likert-type items and Likert-type scales

Likert-type items are similar to Likert items but do not respect some of their characteristics. Despite the lack of agreement on this matter, a good compromise is to follow the proposal of Uebersax [2006] and thus, to define a Likert-type item as a Likert item that is not bivalent and symmetrical about a neutral middle (condition n° 4), and/or does not measure an attitude in terms of level of agreement/disagreement to a target statement (condition n°5). An example of Likert-type item is proposed in Figure 2.



	How much of the time during the past two weeks have you been bothered by	All the time	Most of the time	Slightly more than half the time	Slightly less than half the time	Some of the time	At no time
1	nervousness, tension or inner unrest?	5	4	3	2	1	0
2	worrying too much about even the most insignificant things in your daily life?	5	4	3	2	1	0
3	having to avoid certain things, places or activities as anxiety-provoking?	5	4	3	2	1	0

Figure 3. Exemple of Likert-type scale [Bech 2011]

Clason and Dormody [1994] and Boone, Jr. and Boone [2012] compare Likert scales to Likert-type items. For them the difference between Likert scales and Likert-type items is that the former contains Likert items as single questions that use some aspect of the original Likert response alternatives but do not cluster them into a composite scales. This adds the fact that several items always compose a Likert scale; it is never an individual Likert item. More precisely, a Likert scale is a scale composed only by Likert items. Likert-type items only compose a Likert-type scale. If in a multi-item scale, some of the items are not Likert items, it is more suitable to call this scale a visual analog scale VAS or a Discrete Visual Analog Scale DVAS (presented in the following paragraph).

2.3 Discrete Visual Analog Scales DVAS

A Visual Analog Scale (VAS) is a measurement instrument that tries to measure a characteristic or attitude that is believed to range across a continuum of values. A Discrete Visual Analog Scale is a VAS that provides pre-specified (or discrete) levels to the respondent. Figure 3 visually shows the difference between a VAS and a DVAS. All Likert-type items are DVAS, but not all DVAS are Likert-type items (they may not respect condition n°1, n° 2 or n° 3).

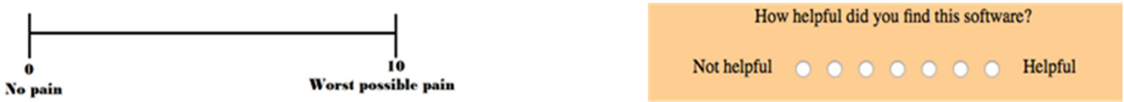

0
No pain
10
Worst possible pain
How helpful did you find this software?
Not helpful ○ ○ ○ ○ ○ ○ ○ Helpful

Figure 4. Example of Visual analog scales (left) and Discrete visual analog scale (right)

2.4 Rating Scales RS

According to *Enciclopedia Britannica*, rating scale is the most general possible term for scales that present users with an item and ask them to select from a number of choices. Rating scales can represent any of a number of concepts. Items can be rated on a single conceptual scale or each may be rated on a series of scales representing a variety of concepts or attributes. Frequently rating scales are combined to create indices [Weller and Romney 1988]

According to Stevens [1946], rating scales are classified in four categories (this classification is also [Blessing and Chakrabarti 2009, pp.118-119]). Other categorizations appear in literature, (e.g. 6 categories in [Guilford 1954]). Stevens 4 categories are:

1. Nominal: a topological scale representing qualitative properties, whose relations can only be defined in terms of equalities. It is the weakest level of measurement representing categories without numerical representation. No statistical analysis can be performed except cross tabulations. Typical representations are bar charts and pie charts.
2. Ordinal: a topological scale representing qualitative properties that can be ranked, but the distance between the categories cannot be said to be equal. An ordering or ranking of responses is possible but no measure of distance is possible. Median, rank correlation, frequency tables, as well as cross tabulations can be applied. Typical representations are bar charts.

3. Interval: a metric scale, where distances between categories are known and equals. Each level of measurement has a relative meaning to each other, but a natural zero is absent. Ordering and distance measurement are possible. Arithmetic mean, standard deviation and related test (e.g. Pearson, F-test, etc.) are applicable. Typical representations are line graphs.
4. Ratio: a metric scale with equal distance between level of measurement and a natural zero. Meaningful ordering, distance, decimals and fractions between variables are possible. All the previous mathematical procedures are possible. Moreover geometric mean, ratio, variance can be used. All type of representations can be used.

The difference between a topological scale and a metric one is that topological scale values are merely ordinal items, while values of metric scale are cardinal [Berka 1983, pp.158-178] (see also http://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/).

According to this scale category a different statistical analysis should be performed. Hence, the next section will debate around the following question: To which category do belong Likert and Likert-type scales?

3. Differences in ordinal and interval scales

The main debate around Likert and Likert-type scales, under a statistical point of view, is whether to consider them as interval or ordinal scales [Jamieson 2004], [Norman 2010]; DVAS are commonly referred as ordinal scales.

This is not a simple ontological debate, but has practical implication. When calculating p-values, which is the level of significance, interval scales should be analyzed with a parametric statistical approach while ordinal scales with a non-parametric one. Table 1 compares those approaches.

Table 1. Nonparametric analog to Parametric Statistical Methods for group comparison [Kuzon et al. 1996], [Blessing and Chakrabarti 2009]

Type of Problem	Type data	Nonparametric Methods	Parametric Methods
Comparison of groups	One group (compared to a reference value)	Chi-squared, Kolmogorov-Smirnov	z-test, t-test
	Two independent groups	Mann-Whitney, Chi-squared, Wilcoxon,'s signed rank, Wald-Wolfowitz, Median test, Kruskall-Wallis, Kolmogorov-Smirnov two-sample	t-test for independent groups, z-test, ANOVA, MANOVA
	Two paired of related groups	Wilcoxon rank sum test, sign test, McNemar's test	Paired T-test, z-test
	Three or more groups	Friedman's two-way ANOVA by ranks, Cochran Q test, Kruskall-Wallis	ANOVA with replication, z-test

The main difference between a non-parametric and a parametric approach consists in the assumption of a normal distribution of the sample population. Knapp [1990] pictures quite well the different position about "conservative" and "liberal" positions about the analysis of Likert scales. Moreover, several insightful discussion and articles are available online on the matter through a simple search.

3.1 Reasons for a non-parametric approach

Several authors state that Likert and Likert-type scales fall within the ordinal level of measurement. [Kuzon et al. 1996], [Jamieson 2004], [Blessing and Chakrabarti 2009], [Boone, Jr. and Boone 2012]. Their position is based on three main topics.

3.1.1 The average of fair and good is not fair and a half

In a rank order, is possible to use discrete values for attitudes or feelings, but you cannot assume that the intervals between these values are equals (i.e. the difference between "agree" and "neutral" cannot be said to be the same difference between "strongly disagree" and "disagree") [Cohen et al. 2000]. Moreover, Jamieson [2004] paraphrasing Kuzon et al. [1996] uses the following example: "*the average of fair and good is not fair and a half even when integers are assigned to represent fair and good*".

3.1.2 We do not have a normal distribution

Data should respect normal distribution for a parametric analysis [Kuzon et al. 1996]. However, as Clason and Dormody defend: " It is difficult to see how normally distributed data can arise in a single Likert-type item, with data frequently skewed, ..." [Clason and Dordomdy 1994].

3.1.3 Size does matter

Kuzon et al. [1996] defend the idea that for small samples a parametric approach is the second sin of statistical analysis. They suggest that a strict attitude toward sample size could be that parametric analysis is appropriate only if N > 10 or even N <30 for each group (N=size of the sample). This idea is reproduced in a lot of website (the faster source of information for students) across the web (e.g. www.researchgate.net/post/What_are_the_specific_rules_for_the_use_parametric_and_nonparametric_tools or http://statsthewayilikeit.com/about/parametric-v-non-parametric-data/).

3.2 Reasons for a parametric approach

On the other hand previous reasons are refuted by "liberals" researchers [Knapp 1990] who defend the use of a parametric approach.

3.2.1 The average of fair and good is fair and a half

Different theses are used to rebut this point. Liberals do not argue again the "fair and a half proof", but they point out that researchers have advantages to ignore the latter evidence and threat ordinal data with a parametric approach. Clason and Dormody [1994] suggest that Likert and scales, "*for their nature and characteristics, or how data are obtained using them, should be analyzed with maximal sensitivity and power.*" Carifio and Perla add "*The debate on Likert scales and how they should be analyzed, therefore, clearly and strongly goes to the intervalist position, if one is analyzing more than a single Likert item. Analyzing a single Likert item, it should also be noted, is a practice that should only occur very rarely… Treating the data from Likert scales as ordinal in character prevents one from using these more sophisticated and powerful modes of analyzes and, as a result, from benefiting from the richer, more powerful and more nuanced understanding they produce* " [Carifio and Perla 2008, p.1151]. Norman in a more recent paper supports another topic, that in his opinion is stronger than the previous two. He defends the idea that parametric test, e.g. Pearson test, are insensitive to extreme violations of the basic assumptions concerning ordinal scales [Norman 2010].

3.2.2 We do not need a normal distribution

Norman [2010] goes against the idea that parametric tests cannot be used if the distribution is not normal: *"...parametric methods examining differences between means, for sample sizes greater than 5, do not require the assumption of normality, and will yield nearly correct answers even for manifestly non-normal and asymmetric distributions like exponentials."*

3.2.3 Size does not matter

Norman also confutes the impact of the sample size: "Nowhere in the assumptions of parametric statistics is there any restriction on sample size. Nor is it the case that below some magical sample size, one should use non-parametric statistics. Nowhere is there any evidence that non-parametric tests are more appropriate than parametric tests when sample sizes get smaller [Norman 2010].

Saying the exact opposite of authors in section 3.1.3, others researchers state that nonparametric tests usually require a larger sample size (n value) to have the same statistical power [Lehmann 1998]

[Sullivan and Artino Jr. 2013]. All those "liberals" reasons are summarized by Norman statement: "*Parametric statistics can be used with Likert data, with small sample sizes, with unequal variances, and with non-normal distributions, with no fear of 'coming to the wrong conclusion'. These findings are consistent with empirical literature dating back nearly 80 years* " [Norman 2010].

3.3 Case Study

Both fields seem to have valid reasons, thus what a novice design researcher (that does not have particularly envy to dig into this breathtaking debate) should do?
In order to show the impact of different approaches, we consider an example based on paradigm evaluation [Buisine et al. 2012], where small groups have been compared while performing the same activity but mediated by different support (control vs. experimental). Groups answer to questionnaires made of Likert-type items (probably the most common case) from 1-7 to evaluate the quality of their experience.
Let consider the following sets of data (they are represented this way to optimize space).

[Likert-type item 1 – Control] = {5, 7, 7, 7, 7, 7, 7, 6, 5, 6};
[Likert-type item 1 – Experimental] = { 5, 2, 4, 6, 2, 5, 5, 5, 6, 5};

[Likert-type item 3 – Control] = {2, 6, 7, 7, 7, 5, 6, 5, 1, 5};
[Likert-type item 3 – Experimental] = { 6, 6, 6, 5, 3, 5, 6, 5, 6, 7};

[Likert-type item 4 – Control] = {2, 1, 2, 1, 1, 1, 1, 1, 1, 2};
[Likert-type item 4 – Experimental] = { 2, 3, 1, 1, 4, 4, 1, 2, 1, 4};

Table 2 summarizes the statistical analysis as if Likert-type items are linear (i.e. parametric approach).

Table 2. Parametric approach

Likert-type	Mean	Variance	Standard deviation	T-Test
Item 1 - Control	6,4	0,711	0,843	0,002
Item 1 - Experimental	4,5	2,055	1,433	
Item 3 - Control	5,1	4,322	2,078	0,598
Item 3 - Experimental	5,5	1,166	1,080	
Item 4 – Control	1,3	0,233	0,483	**0,047**
Item 4 - Experimental	2,3	1,788	1,337	

Table 3. summarizes the statistical analysis as if Likert-type items are ordinal (i.e. non-parametric)

Table 3. Non-parametric approach

Likert-type	Median	Mode	Mann-Whitney U
Item 1 - Control	7	7	0,003
Item 1 - Experimental	5	5	
Item 3 - Control	5,5	7	0,968
Item 3 - Experimental	6	6	
Item 4 – Control	1	1	**0,121**
Item 4 - Experimental	2	1	

If any value item set has hypothetically a normal distribution, median, mean and mode would be equal. This is not the case, which approach is so preferable, supposing that even ordinal data for N>5 [Norman 2010] can be treated as interval data (i.e. do not require assumptions of normality)?

The mean is a more popular approach; it induces the reader to think of something homogenous, normally distributed, of which I do not care about each element but about the result of the mix of each component. This representation is useful to give fast information over a group population or to assemble several Likert-type items. However, it is not a robust statistical tool, being largely influenced by outliers (skewed distributions).

On the other hand median and mode provide a better insight of each element of the sample population as well as its distribution, a non-parametric approach is able to provide a better description of groups where distribution is highly skewed, but working with median is more complicated and less intuitive, especially when you had to compare several Likert-type items.

If we consider [Likert-type item 2 – Control], the distribution is far from being normal, hence a mean of 5,1. Althought this would suggest a Gaussian around the value of 5, this is not the case as a median of 5,5 with a mode of 7 suggests.

Relating to statistical-significance, as Table 2 and Table 3 show, a non-parametric analysis has a greater strictness. Considering [Liketr-type item 3], while the t-test provides a significant result ($p < 0,05$), the Mann-Whitney U test does provide a non-significant one ($p>0,1$). In term of p-value, a non-parametric approach is more conservative.

4. Conclusion

According to Adams et al. [1965, p.100]: "*Nothing is wrong per se in applying any statistical operation to measurements of given scale, but what may be wrong, depending on what is said about the results of these applications, is that the statement about them will not be empirically meaningful or else that it is not scientifically significant.*" Klapp [1990] asks for a truce granting the liberty of choice according to the position of Adams et al. [1965]. Hence, if both approaches are acceptable, what can we do to improve the rigor of design research? A possibility is to perform both approach and confront results, but it is a very time demanding solution.

A pragmatic suggestion is to follow the proposition of Boone, Jr. and Boone [2012] and to use parametric approach with Likert and Likert-type scales (i.e. every time you have to aggregate a consistent number of Likert, Likert-type, and VAS) and non-paramatric with Likert and Likert-type items (and DVAS). This is similar to what suggested by Carifio and Perla [2008], that in turn precise the proposal of Clason and Dormody [1994]: "*The weight of the empirical evidence, therefore, clearly supports the view and position that Likert scales (collections of Likert items) produce interval data , particularly if the scale meets the standard psychometric rule-of-thumb criterion of comprising at least eight reasonably related items.*" [Carifio and Perla 2008, p.1150]. I think this is a good compromise.

Furthermore, a real advancement toward scientific rigor would be to improve the awareness of design researchers in using statistical significance as discriminant between a "good" and a "bad" result. This is a common but incorrect habit, "*the ritual of null hypothesis significance testing* " [Cohen 1994]. Johnson’s « The insignificance of Statistical Significance Testing » [Johnson 1999] well explain the problem related to a incorrect understanding of the use of statistical significance. Anscombe [1956] observes that statistical hypothesis tests are totally irrelevant, and that what is needed are estimates of magnitudes of effects. "*The primary product of a research inquiry is one or more measures of effect size, not P values.*" [Cohen 1990] via [Sullivan and Feinn 2012] "*Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude – not just, does a factor affect people, but how much does it affect them.* " [Kline 2004] via [Sullvan and Feinn 2012] The effect size is the magnitude of difference between groups. It is the main finding of a quantitative study. It can be calculated both for parametric (e.g. Cohen’s d) and non-parametric approaches (Cliff’s delta).

Related to effect size is the sample size. According to Strasak et al. [2007]: "*It is crucial for high quality statistical work, to consider a-priori effect and sample size estimation and to appropriately conduct a statistical power calculation in the planning stage, to make sure that a study is provided with sufficient statistical power to detect treatment effects under observation."* According to Sullivan and Feinn [2012]: *"An estimate of the effect size is often needed before starting the research endeavor, in order to calculate the number of subjects likely to be required to avoid a Type II, or β, error, which is the probability of concluding there is no effect when one actually exists. In other words, you must determine what number* *of subjects in the study will be sufficient to ensure (to a particular degree of certainty) that the study has acceptable statistical power to support the null hypothesis. That is, if no difference is found between the groups, then this is a true finding.*" However, as noted by Mumby [2002], while for parametric statistic " *the power of parametric tests can be calculated easily from formulae, tables and graphs*, *the calculation of power for non-parametric techniques is less straightforward* (i.e. it can be calculated using Monte Carlo simulation methods)." Lehmann [1998] suggests to " *compute the sample size required for a parametric test and add 15%* ".

Different opinions nourish this debate, with respectable reasons from both sides. A compromise based on the trade-off between rigor and pragmatism is necessary, exspecially for novice design researchers. They are confronted to the jungle of different research approaches, in which they try to survive and find a path toward the Eldorado of scientific rigor and reproducibility of results. Conseqently, they can only benefit from beaten paths as standard procedure and research methodologies. In this sense, this contribution humbly follows the spirit of the work of Blessing and Chakrabarti with their tentative to provide a common Design Research Methodology.

The next and final section proposes a common procedure (i.e. guidelines) for novice design researcher when using Likert-type scales in small groups comparative design observations. This procedure is of course perfectible. The main goal of this contribution is to have a real debate on the matter, and that is why I have chosen Design conference, in order to arbitrarily decide as a community which approach we prefer. This will ease the life of researchers who may not be willing to spend hours digging in this statistical debate. As Pearson said "*Statistics is the grammar of science*", but quoting Jeffrey Gitomer: *your grammar is a reflection of your image. Good or bad you have made an impression. And like all impressions, you are in total control"*. Hence, it is up to us to decide how the grammar of statistics should be used to form the language of design research.

5. A proposal common procedure when using Likert (Likert-type and VAS) scales/items in small groups comparative design observations

Table 4 provides practical answer for each type of common task a researcher has to face when using Likert (Likert-type and VAS) scales. All together they form a procedure that can ease the use of likert scale and increase the scientific rigour of the research study.

Table 4. Proposal of a common procedure when using

Type of Problem	Parametric Methods (for Likert, Likert –type and VAS scales)	Nonparametric Methods (for items and DVAS)
Estimate sample size	Through Z-score table, several calculator online (e.g. http://www.uccs.edu/~lbecker/, http://www.raosoft.com/samplesize.html https://www.checkmarket.com/market-research-resources/sample-size-calculator/)	Parametric sample size * 1,15
Estimate Effect size	risk difference, risk ratio, odds ratio, Cohen's *d*, Glass's delta, Hedges' *g*, the probability of superiority (a good guide to choose is: http://www.psychometrica.de/effect_size.html)	Cliff's delta
Estimate Statistical Power	Equal to 1-β. Use Cohen's power tables. A common belief is that minimum acceptable value is 0.8 (80%); with small group you will rarely reach this level of statistical power, even for very strong effects (Cohens'd = 1.3) and high alpha (α=0.05 or even α=0.1).	Monte Carlo simulations
Estimate Statistical significance for independent groups	t-test, z-test, ANOVA.	Mann-Whitney

Estimate Statistical significance for related groups	Paired T-test, z-test.	Wilcoxon rank sum test, sign test

This table can be resumed in the following procedure:

1. Choose wether you will use a parametric or non-parametric approach;
2. Choose the level of confidence (http://www.wikihow.com/Calculate-Confidence-Interval) (common values are 90 – 95 -99);
3. Chose the margin of error - precision (common values 5% - 2,5% -1%);
4. Estimate the sample size;
5. Estimate the effect size;
6. Estimate statistical power;
7. Estimate Statistical significance;
8. Report all these information.

An example of how to report such information: ""*Among 7th graders in Lowndes County Schools taking the CRCT reading exam (N = 336), there was a statistically significant difference between the two teaching teams, team 1 (M = 818.92, SD = 16.11) and team 2 (M = 828.28, SD = 14.09), t(98) = 3.09, p ≤ .05, CI.95 -15.37, -3.35. Therefore, we reject the null hypothesis that there is no difference in reading scores between teaching teams 1 and 2. Further, Cohen's effect size value (d = .62) suggested a moderate to high practical significance* " [Biddix 2009].

References

Adams, E., Fagot, R. F., Robinson, R. E., "A theory of appropriate statistics", Psychometrika, Vol.30, No.2, 1965, pp. 99-127.

Anscombe, F. J., "Discussion on Dr. David's and Dr. Johnson's Paper", J. R. Stat. Soc., Vol.18, 1965, pp. 24-27.

Berka, K., "Measurement: Its Concepts, Theories and Problems", Springer Science & Business Media, 2012.

Biddix, P., "Uncomplicated Reviews of Educational Research Methods", Available at: <https://researchrundowns.wordpress.com/quantitative-methods/effect-size/>, 2009, [Accessed 12.01.2015].

Blessing, L. T. M., Chakrabarti, A., "DRM: Design Research Methodology", Springer-Verlag London, 2009.

Blessing, L. T. M., "A process-based approach to computer-supported engineering design", Univ. of Twente, 1994.

Boone Jr, H. N., Boone, D. A., "Analyzing Likert Data", Journal of Extension, Vol.50, 2012.

Buisine, S., Besacier, G., Aoussat, A., Vernier, F., "How do interactive tabletop systems influence collaboration?", Computers in Human Behavior, Vol.28, No.1, 2012, pp. 49–59.

Carifio, J., Perla, R., "Resolving the 50-Year Debate around Using and Misusing Likert Scales", Medical Education, Vol.42, No.12, 2008, pp. 1150–1152.

Clason, D. L., Dormody, T. J., "Analyzing Data Measured By Individual Likert-Type Items", Journal of Agricultural Education, Vol.35, 1994, pp. 31–35.

Cohen, J., "The Earth Is Round (p < 0.5)", American Psychologist, Vol.49, No.12, 1994, pp. 997–1003.

Cohen, L., Manion, L., Morrison, K., "Research Methods in Education", 5th ed., Routledge Falmer London, 2000.

Gidel, T., Kendira, A., Jones, A., Lenne, D., Barthès, J.-P., Moulin, C., "Conducting Preliminary Design around an Interactive Tabletop", DS 68-2: Proceedings of ICED 11, Vol.2, 2011, pp. 366-376 .

Guerra, A. L., Gidel, T., Vezzetti, E., "Digital intermediary objects: the (currently) unique advantage of computer-supported design tools", DS 80-5: Proceedings of ICED 15, Vol.5 - Part 1, 2015, pp. 265-274.

Guilford, J. P., "Psychometric Methods", Tata McGraw Hill Publishing CO Ltd. New Delhi, 1954.

Hinds, P. J., Kiesler, S. B., "Distributed Work", MIT Press, 2002.

Jamieson, S., "Likert scales: how to (ab)use them", Medical Education, Vol.38, No.12, 2004, pp. 1217–1218.

Johnson, D. H., "The insignificance of statistical significance testing", The Journal of Wildlife Management, Vol.63, No.3, pp. 763-772.

Bales, R. F., "Interaction process analysis; a method for the study of small groups", Addison-Wesley Oxford England, 1950.

Jupp, V., "The SAGE Dictionary of Social Research Methods", SAGE, 2006.

Knapp, T. R., "Treating Ordinal Scales as Interval Scales: an Attempt to tesole the controversy", Nursing Research, Vol.39, No.2, 1990, pp. 121-123.

Kuzon, W. M., Urbanchek, M. G., McCabe, S., "The Seven Deadly Sins of Statistical Analysis", Annals of Plastic Surgery, Vol.37, No.3, 1996, pp. 265–272.

Lehmann, E. L., "Nonparametrics : Statistical Methods Based on Ranks", Prentice-Hall, 1998, pp. 76-81.

Likert, R., "Likert technique for attitude measurement", Social Psychology: Experimentation, Theory, Research, Sahakian, W. S. (Ed.), Intext Educational Publishers Scranton, USA, 1972, pp. 101-119.

Maher, M.-L., Cicognani, A., Simoff, S., "An experimental study of computer mediated collaborative design", Proceedings of the 5th Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, 1996, pp. 268–273.

Mumby, P. J., "Statistical Power of Non-Parametric Tests: A Quick Guide for Designing Sampling Strategies", Marine Pollution Bulletin, Vol.44, No.1, 2002, pp. 85–87.

Norman, G., "Likert Scales, Levels of Measurement and the 'laws' of Statistics", Advances in Health Sciences Education, Vol.15, No.5, 2010, pp. 625–632.

Stevens, S. S., "On the Theory of Scales of Measurement", Science, Vol.103, No.2684, 1946, pp. 677-680.

Strasak, A. M., Zaman, Q., Pfeiffer, K. P., Ulmer, H., "Statistical Errors in Medical Research--a Review of Common Pitfalls", Swiss Medical Weekly, Vol.137, No.3-4, 2007, pp. 44–49.

Sullivan, G. M., Artino Jr., A. R., "Analyzing and Interpreting Data From Likert-Type Scales", Journal of Graduate Medical Education, Vol.5, No.4, 2013, pp. 541-542.

Sullivan, G. M., Feinn, R., "Using Effect Size—or Why the P Value Is Not Enough", Journal of Graduate Medical Education, Vol.4, No.3, 2012, pp. 279–282.

Tang, H. H., Lee, Y. Y., Gero, J. S., "Comparing collaborative co-located and distributed design processes in digital and traditional sketching environments: A protocol study using the function–behaviour–structure coding scheme", Design Studies, Vol.32, 2011, pp. 1–29.

Uebersax, J. S., "Likert scales: dispelling the confusion", Statistical Methods for Rater Agreement website, Available at: <http://john-uebersax.com/stat/likert.htm>, 2006, [Accessed: 12.01.2015.].

Weller, S. C., Romney, A. K., "Systematic Data Collection", SAGE, 1988.

Andrea Luigi Guerra, MSc. Ing.
Sorbonne Universités - Universite de Technologie de Compiegne, Innovation Center - Costech EA 2223
6 Avenue du Marechal Foch, 60200 Compiegne, France
Email: andrea.guerra@utc.fr