116317

The general labelled magnitude scale is more reliable in assessing acute itch intensity than the classic visual analogue scale

Running head: Reliability analysis of itch measurement scales

Number of words: 1100

Number of figures: 1

Number of tables: 2

Olivia Jones, Igor C. Schindler, & Henning Holle

Dept. of Psychology, University of Hull, U.K.

Corresponding author: Dr Henning Holle, Dept. of Psychology, University of Hull, Cottingham Road, Hull, HU6 7RX, U.K., Telephone: +44 (0)1482 466152, Fax: +44 (0)1482 465599, e-mail: h.holle@hull.ac.uk

Funding statement: Parts of this study were supported by a grant from the British Skin Foundation, awarded to HH (project number: 7011s).

The reliable measurement of itch intensity is crucial, both in research as well as clinical contexts. For example, when the reliability of a measurement scale is unknown, it is impossible to determine whether a patient has changed sufficiently to be confident that the change is beyond that which could be attributed to measurement error (1). One factor that might influence the reliability of measurements is the type of rating scale used to assess itch intensity. Previous research (2-4) has documented the retest reliability of different rating scales for assessing chronic itch intensity. However, a retest reliability analysis of rating scales for acute experimental itch, induced using substances such as histamine or cowhage, is currently lacking.

Here, we compare the test-retest reliability of three rating scales commonly used for this purpose. First, we considered the visual analogue scale in its classic form (cVAS), where participants indicate itch intensity on a line ranging from 0 (no itch) to 100 (the most intense itch imaginable). Second, we included a variant of the VAS, where an additional 'Scratch Threshold' marker is set at 33% (tVAS,5), defined as itching strong enough to be scratched (6). Finally, we considered the general Labelled Magnitude Scale (gLMS,7), where participants judge the magnitude of itch on a line with quasi-logarithmically placed labels of "no sensation" at 0, "barely detectable" at 1, "weak" at 6, "moderate" at 17, "strong" at 35, "very strong" at 53 and "strongest imaginable sensation" at 100. Thus, all three scales have an identical range, but differ in the type and number of verbal labels provided.

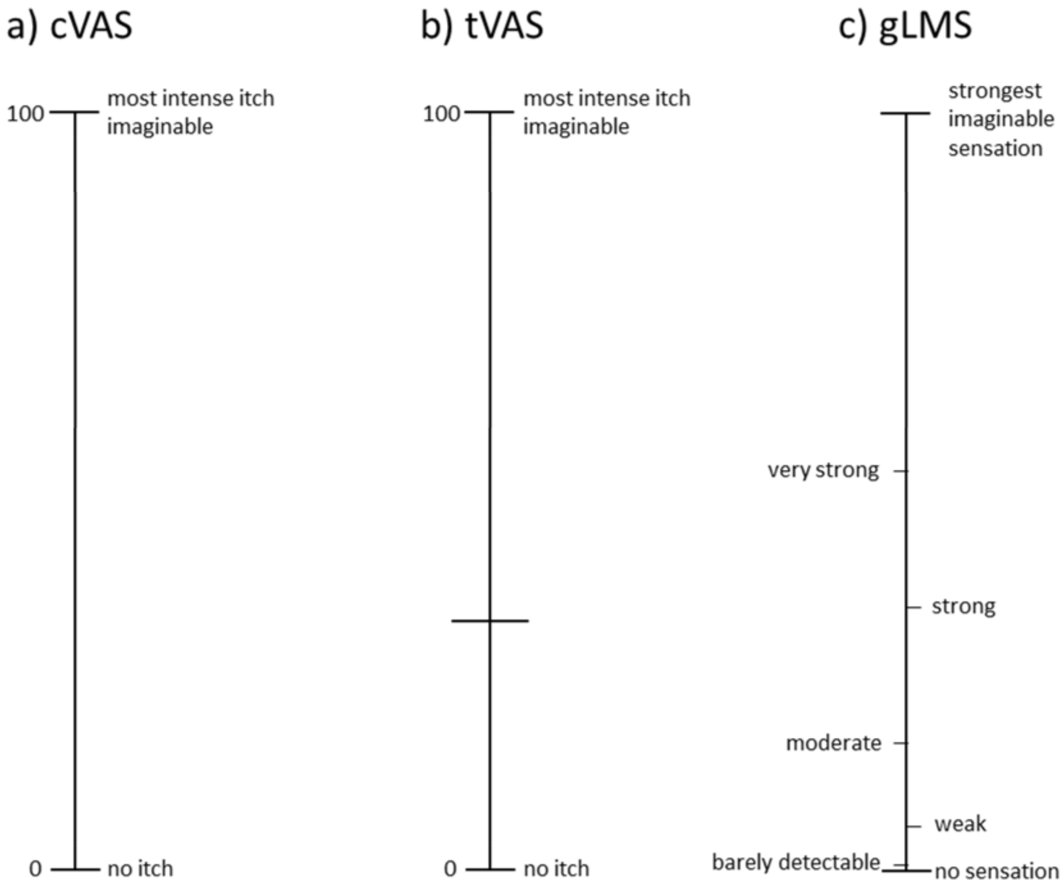


Figure 1 Overview of the three different rating scales used in the present study. Instructions for the classic visual analogue scale (**cVAS**, shown on the left, panel a) were as follows: “The experiment will involve the researcher rubbing spicules from the pod of a cowhage plant into the forearm for 45 seconds. This will feel itchy and may also create a slight stinging/pricking sensation. In each of the sessions you will be required to rate the intensity of the itch every 15 seconds on a computerized scale for 10 minutes. The scale starts from 0 (no itch) to 100 (most intense itch imaginable). Once you have given your response, the scale will disappear until 15 seconds have passed. If you felt that itch intensity fluctuated (went up and down) during the last 15 seconds, you should base your response on the most intense itch perception that you felt during these last 15 seconds.” Instructions for the **tVAS**, centre panel b), were identical except for the following addition. “The additional line at one third of the scale represents the scratch threshold. You should give ratings above this threshold if the itch is so intense that you feel the urge to scratch”. Instructions for the **gLMS** (right panel c) were identical to those of the cVAS, except the wording of the verbal anchors in the instructions was changed accordingly. All participants were reminded verbally that they should not focus on the labelled points of the scale, but could use all points in between as well, depending on the strength of the itch experience.

Ninety healthy volunteers took part after giving written informed consent. Twelve participants (gLMS group: N=7, cVAS group: N=5) were screened out as non-responders after the familiarization session (i.e., itch intensity ratings did not exceed 15) and one as an outlier (itch response above 3 SD of group mean), resulting in a final sample of 77 participants (38 females, mean age 24.66 ±6.5; N=25 in gLMS group, N=26 in cVAS and tVAS group). Participants were told the study was investigating the effect of itch on heart rate and were fully debriefed after the final session. The study was approved by the local Ethics Committee. As an experimental itch model, we used the cowhage provocation paradigm (8). Briefly, 60-65 cowhage spicules were placed into a 16cm² area defined by medical tape on the left volar forearm. Spicules were then rubbed into the skin for 45 seconds. Itch intensity ratings were obtained every 15 seconds for 10 minutes using Presentation Version 17.0 (www.neurobs.com).

Participants were randomly assigned to a scale group (cVAS, tVAS or gLMS) and took part in three experimental sessions (average of 7.04 days, ±1.0 between sessions). Session 1 served as a familiarization session, where participants were trained in the correct application of the rating scale (as recommended by 2) and could experience the novel sensation of cowhage-induced itch.

The peak and mean of each time course were used to quantify the overall itch intensity experienced by a participant. Scores did not differ significantly between sessions (Table 1). Shapiro-Wilk tests indicated that mean and peak scores were normally distributed (all $W > 0.93$, all $p > 0.09$). Scale reliability was estimated by the Intraclass-Correlation-Coefficient (ICC) of the respective scores of Sessions 2 and 3, when participants were familiar with the experience of cowhage-induced itch and the scale. For this retest reliability analysis, we used a two-way mixed model, focusing on absolute agreement between sessions (9).

Table 1 Descriptive statistics of the two itch indices (Mean, Peak) for each session and scale group. Columns 5 and 6 provide the t and p value of an independent samples t-test comparing Sessions 2 and 3.

Scale Group	Index	*M* (*SD*) Session 2	*M* (*SD*) Session 3	*t*	*p*
cVAS (n=26)	Mean	24.38 (13.34)	27.45 (13.60)	1.11	0.28
	Peak	64.92 (20.99)	66.31 (22.08)	0.32	0.75
tVAS (n=26)	Mean	25.19 (12.05)	27.64 (13.78)	1.14	0.26
	Peak	64.04 (21.20)	67.31 (24.54)	0.99	0.33
gLMS (n=25)	Mean	16.24 (7.38)	18.59 (9.74)	1.83	0.08
	Peak	48.12 (21.64)	49.68 (23.68)	0.65	0.52

As shown in Table 2, the gLMS had the highest retest reliability. This was the case regardless of which index was used to quantify itch intensity (peak: ICC=.86; mean: ICC=.71). The cVAS was the least reliable scale (peak: ICC=.50; mean: ICC=.45) and the tVAS had an intermediate reliability (peak: ICC=.73; mean: ICC=.64). Associated *p* values, obtained using Fisher's r-to-Z transformation, indicated that the gLMS was significantly more reliable than the cVAS (*p*=.01, see Table 2).

Table 2: Retest reliability (as estimated by the Intraclass Correlation Coefficient, ICC) for the 3 scales and 95% Confidence Interval

Index	Scale	ICC	95% CI
Mean	cVAS^{a,b}	.45	.09 – .71
	tVAS^{c}	.64	.35 – .82
	gLMS	.71	.44 – .86
Peak	cVAS^{d,e}	.50	.14 – .74

	tVAS[f]	.73	.48 – .87
	gLMS	.86	.72 – .94

p values of pairwise comparisons. **Mean**: (a): cVAS vs. tVAS, *p*=.35; (b): cVAS vs. gLMS, *p*=.18; (c): tVAS vs. gLMS, *p*=.69. **Peak**: (d): cVAS vs. tVAS, *p*=.20; (e): cVAS vs. gLMS, *p*=.01; (f): tVAS vs. gLMS, *p*=.20.

The higher retest reliability of the gLMS cannot be explained in terms of response clustering (i.e., the clustering of ratings around the verbal labels, see Supplementary Online Results). Instead, our data suggest that retest reliability may be linked to the degree to which scales are open to interpretation. Previous research has highlighted that the lack of verbal anchors in the cVAS creates ambiguity, because participants are unsure where exactly they should place their mark (10, 11). This unsystematic variation may limit the reliability of the cVAS. In contrast, the tVAS adds a scratch threshold marker, providing participants with an additional landmark to guide their ratings which increases scale reliability. Finally, the gLMS with its seven verbal anchors is least ambiguous and was found to be the most reliable scale for measuring acute itch.

Another factor that could explain the observed superior reliability of the gLMS is that this scale has been explicitly designed to yield ratio data, whereas it is strongly debated whether the cVAS provides ratio (12) or merely ordinal level data (for review, see 11). There is evidence that rather than providing a linear transformation of the internal representation of stimulus intensity, the cVAS provides only a non-linear representation, with a compression of scores especially at the top end of the scale (10). In contrast, the roughly logarithmic distance between the verbal anchors in the gLMS, determined in a semantic scaling procedure, has been demonstrated to yield ratio level data for ratings of oral sensations (13, 14) though a validation in the domain of itch is still outstanding.

A limitation of the present study is that participants were excluded from taking part in sessions two and three when their intensity ratings did not exceed 15 in the initial familiarization session. No participant in the tVAS group was excluded based on this criterion, but several in the gLMS (N=7) and cVAS (N=5) group, which may have biased the results. In general, obtaining very low ratings seems less likely when using the tVAS. Note, however, that this potential bias cannot explain the main finding of our study (gLMS is significantly more reliable in assessing peak itch than cVAS), since a comparable number of participants were excluded from these two groups.

In summary, our results suggest that the gLMS rating scale enables a more reliable measurement of acute itch intensity in healthy volunteers. The gLMS scale may be particularly suited for longitudinal studies, though care must be taken to avoid memory effects (e.g., by allowing for sufficient time between ratings, or by using distractor items). Since scale reliability is not a fixed property, but is also population-dependent (15), further studies are necessary to investigate whether these advantages of the gLMS scale generalise to experimental itch induced in chronic itch patients or to the clinical assessment of chronic itch intensity.

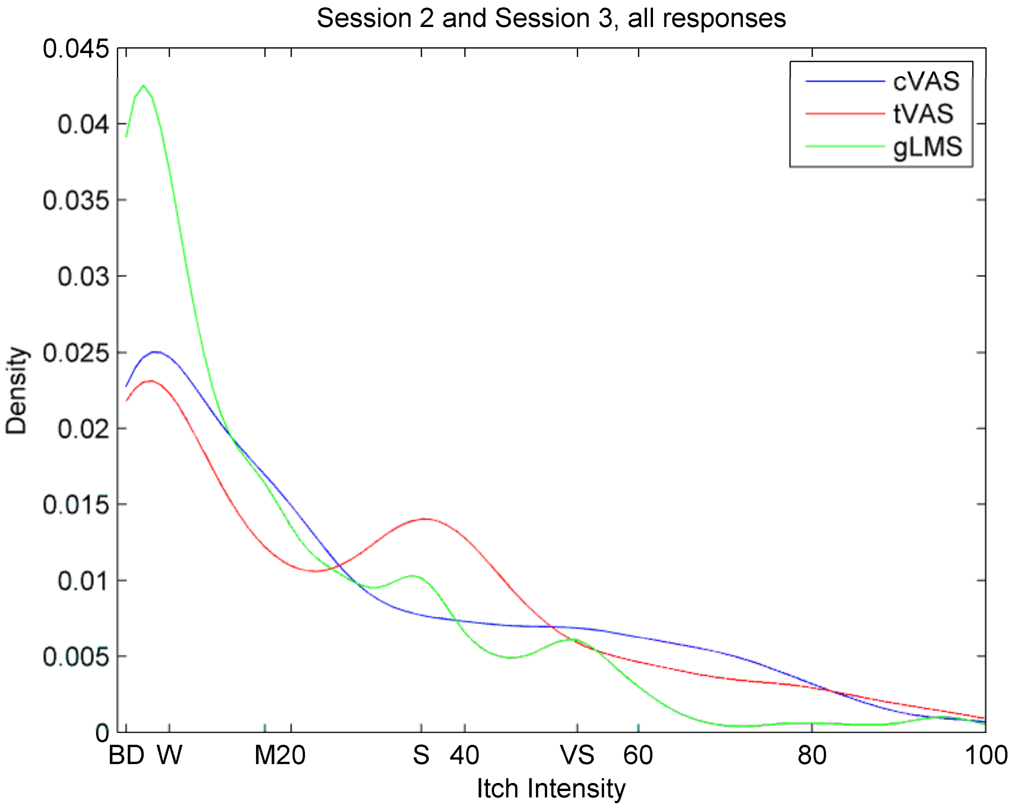
References

1. Evans A, Margison F, Barkham M. The contribution of reliable and clinically significant change methods to evidence-based mental health. Evid Based Ment Health 1998;1:70-72.
2. Phan NQ, Blome C, Fritz F, Gerss J, Reich A, Ebata T, et al. Assessment of pruritus intensity: prospective study on validity and reliability of the visual analogue scale, numerical rating scale and verbal rating scale in 471 patients with chronic pruritus. Acta Derm Venereol 2012;92:502-507.
3. Reich A, Szepietowski JC. Pruritus intensity assessment: challenge for clinicians. Expert Rev Dermatol 2013;8:291-299.
4. Elman S, Hynan LS, Gabriel V, Mayo MJ. The 5-D itch scale: a new measure of pruritus. Br J Dermatol 2010;162:587-593.
5. Darsow U, Ring J, Scharein E, Bromm B. Correlations between histamine-induced wheal, flare and itch. Arch Dermatol Res 1996;288:436-441.
6. Magerl W, Westerman RA, Mohner B, Handwerker HO. Properties of transdermal histamine iontophoresis: differential effects of season, gender, and body region. J Invest Dermatol 1990;94:347-352.
7. LaMotte RH, Shimada SG, Green BG, Zelterman D. Pruritic and nociceptive sensations and dysesthesias from a spicule of cowhage. J Neurophysiol 2009;101:1430-1443.
8. Papoiu AD, Tey HL, Coghill RC, Wang H, Yosipovitch G. Cowhage-induced itch as an experimental model for pruritus. A comparative study with histamine-induced itch. PLoS One 2011;6:e17786.
9. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1:30-46.
10. Gonzalez-Fernandez M, Ghosh N, Ellison T, McLeod JC, Pelletier CA, Williams K. Moving beyond the limitations of the visual analog scale for measuring pain: novel use of the general labeled magnitude scale in a clinical setting. Am J Phys Med Rehabil 2014;93:75-81.
11. Kersten P, Kucukdeveci AA, Tennant A. The use of the Visual Analogue Scale (VAS) in rehabilitation outcomes. J Rehabil Med 2012;44:609-610.
12. Price DD, McGrath PA, Rafii A, Buckingham B. The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. Pain 1983;17:45-56.
13. Green BG, Dalton P, Cowart B, Shaffer G, Rankin K, Higgins J. Evaluating the ‘Labeled Magnitude Scale’ for Measuring Sensations of Taste and Smell. Chem Senses 1996;21:323-334.
14. Green BG, Shaffer GS, Gilmore MM. Derivation and evaluation of a semantic scale of oral sensation magnitude with apparent ratio properties. Chem Senses 1993;18:683-702.
15. Shrout PE. Measurement reliability and agreement in psychiatry. Stat Methods Med Res 1998;7:301-317.

Supplementary Online Results

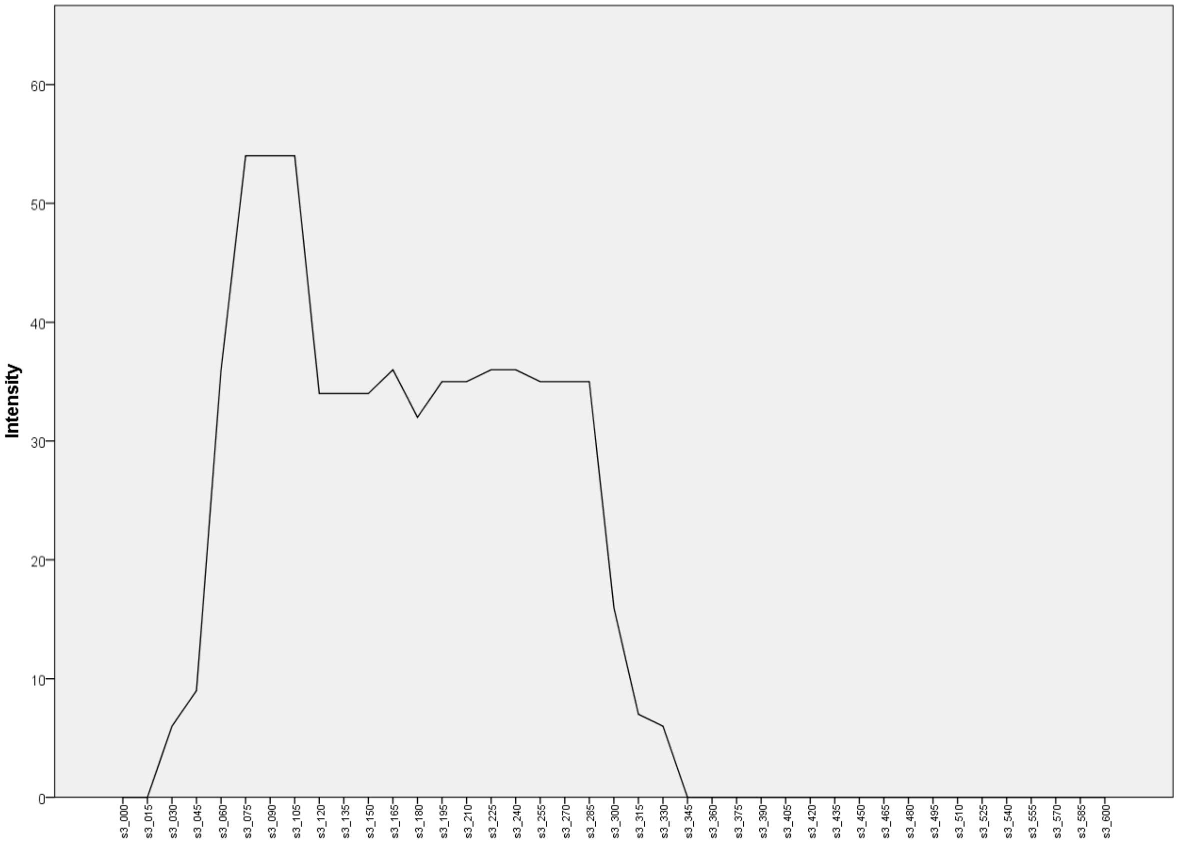
One possible explanation of the superior retest reliability of the gLMS could be that participants may cluster their responses only around the labelled adjectives. This may effectively restrict the spread of responses in the gLMS (which has seven labelled adjectives), but less so in the cVAS and tVAS group, which has fewer labelled adjectives. Such a categorical use of the gLMS has been observed before in the domain of taste perception (16).

As can be seen in Supplementary Figure 1, there is only little evidence of categorical rating behaviour in the gLMS group (especially when comparing it with the strength of previously observed categorical behaviour, see Hayes et al, 2013, Fig. 3). There are no discernible peaks around the labelled positions for 'barely detectable', 'weak', and 'moderate', but some evidence of clustering of responses around the labelled positions for 'strong' and 'very strong' positions.



Supplementary Figure 1 Kernel density plot of all 6314 responses (77 participants * 2 Sessions * 41 time points), separately for each scale group. Labelled points of the gLMS are shown at the bottom of the graph. Abbreviations and position on the 0-100 scale of the labels are as follows: BD (1), Barely detectable; W (6), Weak; M (17), Moderate; S (35), Strong; VS (53), Very Strong.

To further analyse this issue, we looked at the rating timecourses of individual participants from the gLMS group and found that 2 out of 25 participants were indeed using the scale in a more categorical way, rather than (as instructed) in a continuous fashion (see Supplementary Figure 2 for an example timecourse). If retest reliability of the gLMS were largely driven by the presence of categorical rating behaviour, then excluding these two subjects should result in a marked reduction of reliability. As reported in the main paper, the reliability indices (ICC) of the gLMS for the full sample, n=25, are .86 and, .71, for peak and mean, respectively. When excluding the two above-mentioned participants exhibiting categorical rating behaviour, these indices are .87, and .72, respectively. Thus, categorical use of the gLMS occurred only in 2 out of 25 participants, and its presence does not influence scale reliability.



Supplementary Figure 2 Example of a rating timecourse from a single subject from the gLMS group exhibiting categorical use of the scale

Additional References

16. Hayes JE, Allen AL, Bennett SM. Direct comparison of the generalized Visual Analog Scale (gVAS) and general Labeled Magnitude Scale (gLMS). Food Qual Prefer 2013;28:36-44.