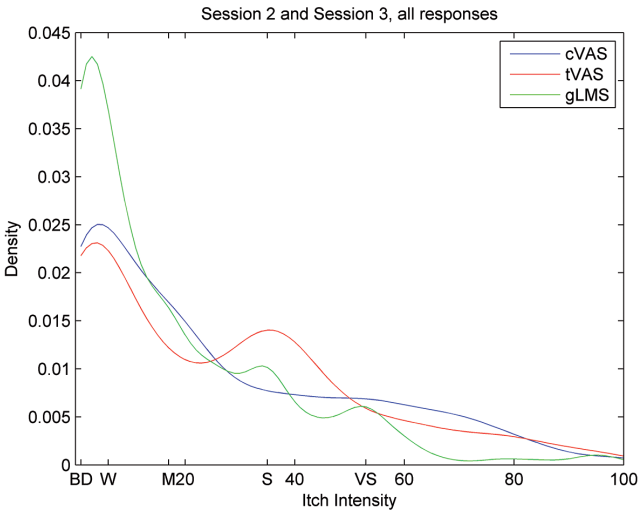
**APPENDIX S1**

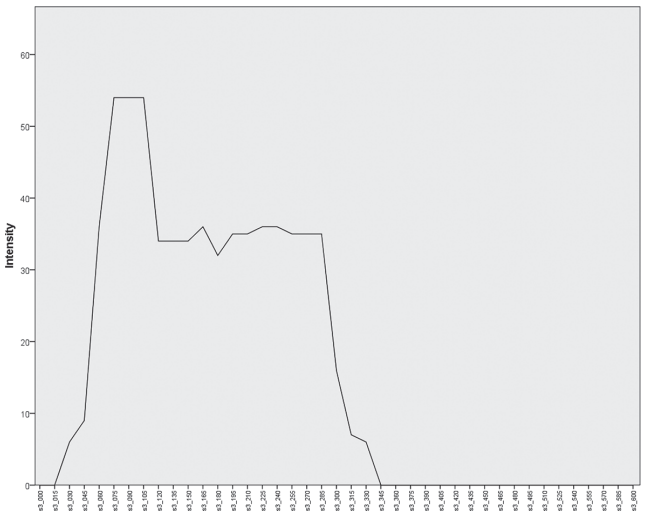
## SUPPLEMENTARY RESULTS

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 7 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 (S1).

As can be seen in **SFig. 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, **SFig. 2**). 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.




**SFig. 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.




**SFig. 2. Example of a rating timecourse from a single subject from the gLMS group exhibiting categorical use of the scale.**

To further analyse this issue, we looked at the rating time courses 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 SFig. 2 for an example time course). 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 0.86 and, 0.71, for peak and mean, respectively. When excluding the two above-mentioned participants exhibiting categorical rating behaviour, these indices are 0.87, and 0.72, respectively. Thus, categorical use of the gLMS occurred only in 2 out of 25 participants, and its presence does not influence scale reliability.

## SUPPLEMENTARY REFERENCE

S1. 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.