Title: Investigation of conceptual distinction between response set options - a step toward obtaining quality data

Authors: Tim Poepsel, PhD (Survey Research Analyst Team Lead), Rachna Kaul, MPA, MDS (Instrument Consultant, Medidata Solutions), Allyson Nolde, MPP (Survey Research Analyst), Rebecca Israel, MS (Survey Research Analyst), Hayley Simpson-Finch, (Director of Accounts, Linguistic Validation), Rachael Browning, BA, (Linguistic Validation Services Manager), Elizabeth McCullough, MA (Senior Director of LV Operations), Mrunmayee Pitkar, MA (Director of Project Management), Shawn McKown, MA (Vice President of Clinical Outcome Assessments)

Introduction:

Ordinal scales with discrete response options are commonly used in assessment of patient's experience with a given health condition or treatment. Close conceptual spacing within response sets can interfere with translation during linguistic validation (LV), reduce comprehension for patients, and impact data validity. Traditional LV makes qualitative assessments of translations via cognitive debriefing to generate quality translations that are faithful to source content, but these qualitative assessments can't diagnose or remedy poorly designed response sets. The present research leverages patient ratings of commonly used response options to quantitatively assess and visualize their distribution in conceptual space, and to provide data to aid in the selection of conceptually distinct response options.

Methods:

115 English-speaking participants from 12 locales (see Table 1) rated between 1 and 3 response options from a set of 15 (see Table 2) on a VAS scale, with 0 representing a minimum severity rating and 100 representing a maximum. The number of ratings for different response options collected from each participant was limited to decrease any possible biasing or anchoring effects caused by making consecutive and conceptually related ratings.

Participants responded to prompts such as those in Figure 1 by moving a slider along a horizontal line. Participants received no instructions regarding the task other than the description provided in the prompt itself. The rating task was not presented with the context typically provided on a VAS scale; participants only rated each response option based on where they believed it should appear between the scale ends of "lowest level of severity" and "highest level of severity". On a typical VAS scale, a rating would be attached to an evaluation of the severity of one or more of a patient's symptoms (e.g., "Please rate the severity of your [SYMPTOM/S] over the past week.").

	Locale	Number of Participants (English-Speaking)
1	Australia	15
2	Belgium	5
3	Canada	15
4	Hong Kong	10
5	India	10
6	Ireland	10
7	Israel	5
8	Malaysia	5
9	New Zealand	10
10	Singapore	5
11	UK	15
12	South Africa	10

Table 1 Data was collected from English-speaking participants from 12 locales

1. Please indicate the level of symptom severity that you think the term **"extremely severe"** represents by placing the slider somewhere along the horizontal line between 0 (lowest level of severity) and 100 (highest level of severity).

0 - (Lowest level of severity)   (Highest level of severity) - 100

Figure 1 An example of the prompt that participants responded to in rating response options

Results:

Figure 2 shows rating distributions for the 15 response options. Patient ratings generally matched expectations, although variability within all rating distributions was unexpectedly high. Distributions for low severity response options 'barely', 'a little' and 'mildly' were clustered and highly variable. Distributions for high severity response options were also

clustered and highly variable; both results suggest the presence of conceptual overlap of response options and possible participant confusion. 'Moderately' and 'somewhat', typical scale midpoints, showed different rating distributions, most notably in terms of variability, suggesting better fit for 'moderately' compared to "somewhat" as a scale midpoint.

Response options "very severe", "quite a bit severe", "severe", and "quite severe" were also clustered in their ratings, and highly variable, suggesting roughly similar interpretations, and again indicating a possible lack of precision regarding intended magnitude. We note another clustered set of response options toward the high severity end of the scale. "Markedly severe", "worst imaginable severity", "most severe", and "extremely severe" appear to be interpreted similarly by participants.

	Response Option	Average VAS Rating	Median VAS Rating	Minimum VAS Rating	Maximum VAS Rating	Mode	Number of Ratings by Participants
1	not at all severe	7.8	0	0	43	0	12
2	hardly severe	31.2	35	1	62	10	13
3	barely severe	25.2	16	0	76	40	15
4	a little severe	41	32.5	12	81	12	14
5	mildly severe	38	32.5	10	73	50	10
6	moderately severe	52	50	30	70	50	11
7	somewhat severe	57.4	70	19	80	80	8
8	quite a bit severe	69.5	68	45	92	80	10
9	severe	67.1	74	25	90	75	14
10	quite severe	69.6	75	13	90	80	12
11	very severe	81.5	82	70	90	90	10
12	markedly severe	86.9	95	60	100	100	8
13	extremely severe	95.3	98.5	75	100	100	12
14	most severe	89	100	50	100	100	10

15	worst imaginable severity	94.5	99	80	100	100	4

Table 2 Descriptive statistics for the data collected on fifteen common response options for severity scales
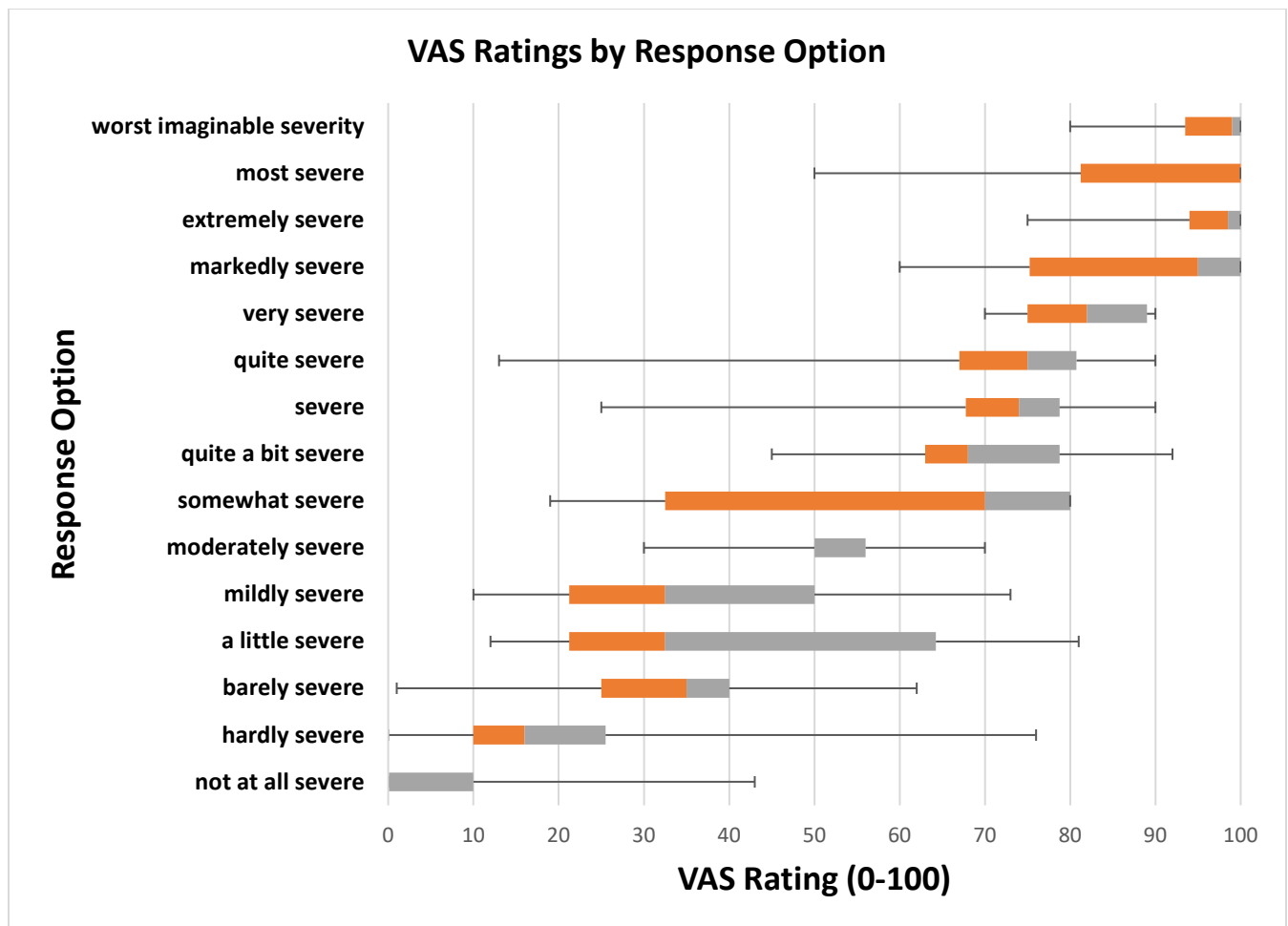
Figure 2 Rating distributions for each response option tested. A rating of 0 indicates "lowest level of severity", while a rating of 100 indicates "highest level of severity".

Conclusion:

These results indicate conceptual clustering of response options that are intended to be distinct, as well as highly variable rating distributions possibly demonstrating poor target concept comprehension. Increased attention to how response sets function internally, and how patients, as opposed to Clinical Outcomes Assessment (COA) professionals, interpret individual response options, can positively affect data validity while improving the patient experience of responding to COAs.

Limitations include different amounts of data collected for each response option, as well as possible dialectical differences between English-speaking locales which may drive some of the distributional variability evident in the findings. Ongoing data collection will address the

low observations for some response options, and also provide opportunity for direct comparisons of rating distributions across English-speaking locales. These findings and the planned future work may provide insight and guidance in selecting maximally distinct response sets, or for conducting post-hoc analyses of data collected via large and possibly problematic response sets