Initial and revised (in parentheses after symbol =) score values of test items (k11-k16, k21-k26, v11-v16, v21-v26) based on the ability mean (AM) of all options of the items used in IRT analysis and explanations and descriptions of actions taken for revising scoring schema if needed.

Item	Data code*	Score value**	Freq.* **(%)	AM* ***	SD of AM	Action taken
k11	6	1	7	.58	.17	No changes were needed. Experts' estimate was 1 and it was therefore expected that only a few respondents would select high values. However, quite a few respondents still chose the worst option as the best. This suggests that in addition to information which makes this option a bad choice, there might still be too much information which makes respondents (presumably beginners) consider it a good choice.
	5	2	6	.78	.29	
	4	3	10	1.04	.26	
	3	4	20	1.13	.36	
	2	5	24	1.35	.49	
	1	6	35	1.58	.58	
k12	6	2	3	.55	.36	No changes were needed. Although there were very few respondents who selected options 6 and 5, their ability mean was significantly different from others. Experts' estimate was 2 and, again, it was expected that only a few respondents would select high values. Compared to item k11, it seems, though, that this item was less likely to be considered a good option even though it was closer to them. Substantive reasoning might be that in the case of this item, the test developers included less information, which made respondents consider this option a good or a very good option.
	5	3	3	.71	.37	
	4	4	6	1.01	.45	
	1	5	38	1.23	.43	
	3	5	21	1.35	.55	
	2	6	29	1.47	.59	
k13	6	3	2	.91	.25	Data codes 6 and 5 were merged because of a similar ability mean and low number of answers for both codes; the new score value for these is 3 points (the ability mean difference from the data score awarded with 4 points is about the same as for the data score awarded with 5 points). Experts' estimate was 3 and thus respondents were expected to select 2–4. Choices 5 and 6 had both very few selections and, from a substantive point of view, it might be that there is just enough information in this item for it to be considered close to expert's estimate. Thus, merging seems reasonable also from a substantive point of view.
	5	4=3	5	.92	.47	
	1	4	18	1.08	.43	
	4	5	12	1.20	.61	
	2	5	30	1.21	.46	
	3	6	33	1.54	.56	
k14	1	3=2	6	.71	.30	Data codes 2, 5 and 3 were initially merged because of a similar ability mean; the new score value for these was planned to be 4 points, as the ability mean difference from the most demanding option (code 4) was significantly larger than this in comparison with the code that was one level below (code 6); subsequently, codes 6 and 1 had to be awarded 1 point less. However, in the content analysis it appeared that the characteristics of option 2 were clearly showing that it was worse than options 3 and 5; thus, it would be reasonable to merge it with option 6 and give them both 3 points, although the ability mean of option 2 is higher than for options 3 and 5.
	6	4=3	8	.95	.38	
	2	4=3	12	1.18	.51	
	5	5=4	16	1.07	.39	
	3	5=4	15	1.14	.39	
	4	6	42	1.57	.55	
k15	1	2	2	.66	.30	Data code 4 had the same score as data code 6, but the ability mean was lower and the score was therefore lowered; subsequently, the other lower scores were lowered as well, but the score of data code 1 was not lowered because it was merged with data code 2, as in both of these cases, very few respondents were selecting these options. Experts' estimate was 5; therefore,
	2	3=2	4	.49	.31	
	3	4=3	6	.90	.41	
	4	5=4	17	1.02	.40	

	6	5	30	1.28	.42	selections 4–6 were expected. From a substantive point of view, choice 1 and 2 are clearly both very bad, so 2 points for both seems to be justified. In the case of choice 4, lowering the points is also justified because choice 6 is clearly better and thus worth more points.
	5	6	41	1.54	.56	
k16	1	1=2	1	.80	.13	Data codes 1, 2 and 3 were merged because of a similar low ability mean and a small number of respondents selecting those options; the new score is 2 because the ability mean differs more from the 5 point code than the 5 point code differs from the 4 point code. Experts' estimate was 6 and it was therefore not expected that very many respondents would choose options 1–3. Even though option 3 seems to be a little bit better than 1 and 2, it might still be reasonable, from a substantive point of view, to merge them all.
	2	2	2	.67	.27	
	3	3=2	5	.54	.34	
	4	4	13	1.01	.32	
	5	5	29	1.23	.47	
	6	6	49	1.48	.54	
k21	6	1	5	.53	.19	Data codes 4, 3 and 2 were merged because of a similar ability mean; the new score value for these is 4 points (their average). Experts' estimate was 1 and it was therefore expected that only a few respondents would select high values. However, from a substantive point of view, choices 5 and 6 are clearly not poor options and should therefore both get 2 points. Choices 3 and 4 seem also quite equal and should therefore both get 3 points. Choice 2 is clearly a worse option than choice 3 and should therefore get 4 points.
	5	2	12	.84	.28	
	4	3=4	10	1.16	.40	
	3	4	16	1.18	.28	
	2	5=4	17	1.21	.32	
	1	6	39	1.60	.61	
k22	6	2=1	1	.34	.24	The ability mean for data code 2 is significantly more different than the differences between other data codes; therefore, the score of the other ones was lowered; codes 6, 5 and 4 were selected by very few respondents, but their ability means are too different for merging these options. Experts' estimate was 2 and responses 5, 6 and even 4 were therefore not expected to appear often.
	5	3=2	2	.76	.23	
	4	4=3	5	.96	.38	
	1	5=4	38	1.13	.35	
	3	5=4	16	1.22	.53	
	2	6	37	1.54	.61	
k23	6	3=1	1	.65	.14	The ability mean for data code 4 was significantly higher than for data code 2, but the scores were the same; similarly, for data code 5 the ability mean was higher than for data code 1; in order to solve this issue, the score of data codes 2 and 1 were lowered by 1 point and, subsequently, the lower scores were lowered as well. Experts' estimate was 3 and responses between 2 and 4 were therefore expected. However, it seems that more respondents preferred choice 2, which is reasonable because choice 3 includes information which clearly makes it a bad choice. As choice 5 is clearly distinguishable as not a good option, then, from a substantive point of view, it should also get less points; thus, 2 points seems more reasonable.
	1	4=2	19	.95	.39	
	5	4=3	3	1.07	.46	
	2	5=4	31	1.18	.46	
	4	5	15	1.33	.47	
	3	6	31	1.57	.60	
k24	1	3	2	.85	.35	No changes were needed.
	6	4	12	1.03	.35	
	2	4	9	1.03	.30	
	3	5	30	1.14	.47	
	5	5	18	1.24	.54	
	4	6	29	1.63	.57	

k25	1	2=N/A	1	1.24	.17	Data code 1 was selected by very few respondents, but these appeared to be the ones with a comparatively good total test score; therefore, this option showed an anomaly and was removed from the analysis; the ability mean of data code 4 was significantly lower than that of data code 6, but the scores were the same; therefore, the score for data code 4 was lowered and, subsequently, other lower scores were lowered as well. Experts' estimate was 5; thus, selections 4–6 were expected to appear more often. From a substantive point of view, there is no good explanation as to why the overall better respondents have chosen this option as the worst. Thus, re-scoring seems reasonable.
	2	3=2	4	.46	.23	
	3	4=3	5	.80	.37	
	4	5=4	33	1.08	.47	
	6	5	10	1.30	.36	
	5	6	47	1.51	.53	
k26	1	1=2	0	.51	.00	Data codes 1, 2 and 3 were merged because they had a very small number of respondents and a comparatively similar ability mean value; the new score is 2, as the ability means are more different from the data code with score 4 than the difference in the case of data code with score 5. Experts' estimate was 6 and it was therefore not expected that very many respondents would choose options 1–3. Even though option 1 is clearly the worst one and distinguishable from 2 and 3, it might still be reasonable, from a substantive point of view, to merge them all.
	2	2	2	.83	.42	
	3	3=2	1	.62	.43	
	4	4	7	.98	.42	
	5	5	18	1.06	.43	
	6	6	71	1.39	.54	
v11	6	1=2	2	.84	.40	Data codes 6 and 5 were merged because of a similar ability mean and low number of answers for code 6; the new score value for these is 2 points (the ability mean difference from the data score awarded with 3 points is about the same as for the data score awarded with 4 points). Experts' estimate was 1 and it was therefore expected that only a few respondents would select values 5 and 6. Thus, from a substantive point of view, re-scoring seems reasonable.
	5	2	7	.92	.39	
	4	3	10	1.01	.43	
	3	4	14	1.18	.44	
	2	5	28	1.24	.52	
	1	6	29	1.56	.56	
v12	6	2	1	.65	.64	The score of code 1 was lowered because it was the same as that of code 1, but its ability mean was lower; subsequently, the score of codes 4 and 5 was lowered as well and code 5 was merged with code 6, as there were very few respondents for both of these options. Experts' estimate was 2 and responses 5 and 6 and even 4 were therefore not expected to appear very often. Other than that, from a substantive point of view, re-coding seems reasonable.
	5	3=2	1	.38	.19	
	4	4=3	4	.92	.23	
	1	5=4	61	1.19	.49	
	3	5	9	1.35	.46	
	2	6	24	1.59	.57	
v13	6	3	22	1.04	.40	Data codes 1, 5 and 2 were merged because of a similar ability mean and low number of answers for code 1; the new score value for these is 4 points (a value between those options that require less ability to answer correctly (code 6) and those that require more ability to answer correctly (code 4). Experts' estimate was 3 and responses between 2 and 4 were therefore expected. It is strange, however, that so many respondents considered this option as the best, although it seems that this option includes pieces of information that are not so correct but might seem reasonable to beginners. Other than that, re-scoring seems reasonable from a substantive point of view.
	1	4	5	1.14	.36	
	5	4	15	1.24	.54	
	2	5=4	15	1.19	.52	
	4	5	20	1.34	.41	
	3	6	22	1.55	.69	
v14	1	3	2	.91	.36	The ability means for data codes 3 and 4 were almost the same; therefore, both of these were awarded 6 points (not 5 points, as the ability mean is higher than for code 5, which was awarded 5 points). Experts' estimate was 4; thus, respondents' selections 3–5 were expected and considered acceptable from a substantive
	2	4	10	1.05	.37	
	6	4	17	1.19	.42	

	5	5	21	1.27	.57	point of view. However, choices 3 and 4 seem quite similar and therefore, from a substantive point of view, re-scoring seems reasonable even though this is the first case where the same amount of points was awarded to an option which is not the option suggested by experts.
	3	5=6	21	1.35	.53	
	4	6	29	1.37	.62	
v15	1	2	1	.03	.52	Data codes 3 and 6 were merged because of similar ability means; the new score value for these is 4 points (a value between those options that require less ability to answer correctly (code 2) and those that require more ability to answer correctly (code 4); there were very few answers for option 1, but it cannot be merged with other codes because of a significant difference in the ability mean. Experts' estimate was 5 and selections 4–6 were therefore expected to appear more often. However, option 3 seems to have almost the same amount of selections, which is strange from a substantive point of view. Other than that, re-scoring seems reasonable.
	2	3	9	1.07	.47	
	3	4	23	1.13	.36	
	6	5=4	21	1.16	.51	
	4	5	21	1.39	.62	
	5	6	24	1.50	.57	
v16	1	1	2	.55	.53	No changes were made. There was only a small number of respondents who selected options 1 and 2, but the ability means for these were significantly different; therefore, it was not reasonable to re-score these options. Experts' estimate was 6 and it was therefore not expected that very many respondents would choose options 1–3. Thus, from a substantive point of view, this item does not need re-scoring either.
	2	2	3	.86	.24	
	3	3	12	.94	.43	
	4	4	16	1.10	.42	
	5	5	31	1.24	.45	
	6	6	37	1.55	.57	
v21	6	1	0	.01	.00	Data codes 4, 3 and 2 were merged because of similar values of ability means and low number of respondents for codes 4 and 3; the new score is 5 because the ability mean is not very different from code 1 with a score of 6. Experts' estimate was 1 and selections 1–3 were therefore expected. Interestingly, in this case, number 1 was chosen overwhelmingly and, from a substantive point of view, the reason seems to be one piece of information inside this item (Motorola case). However, cases 3 and 4 are still clearly worse compared to option 2; thus, from a substantive point of view, awarding 4 points to those options and 5 points to option 2 is justified.
	4	3=4	0	1.17	.00	
	3	4	2	.93	.35	
	2	5	16	1.06	.42	
	1	6	81	1.33	.55	
v22	5	3	1	.61	.48	Data code 4 was scored 1 point higher, as it had an ability mean value comparatively similar to data codes 3 and 1 and it was more different from data code 5, which was scored 3 points, than from data code 2, which was scored 6 points. Experts' estimate was 2 and selections 1–3 were therefore expected. In this case, option 2 was chosen most often and this, again, indicates that there is some information inside it which clearly makes it easy to distinguish as second worse option. In this case, option 6 was never selected, which confirms that overall, situation D2 is easier for respondents due to easily distinguishable items. From a substantive point of view, however, this re-scoring seems justified.
	4	4=5	3	1.01	.53	
	3	5	15	1.06	.39	
	1	5	13	1.11	.45	
	2	6	67	1.38	.56	
v23	6	3	5	.87	.45	Data codes 5 and 4 were merged because of similar ability means; the new score value for these is 4 points, as the ability mean is lower than that of code 2, where it is 5. Subsequently, the score of data code 1 should be lowered and merged with data code 6; merging is appropriate due to a low number of respondents selecting this option. Experts' estimate was 3 and selections 2–4
	1	4=3	3	1.04	.27	
	5	4	18	1.17	.54	
	4	5=4	25	1.19	.40	

	2	5	7	1.26	.42	were expected. Choice 3 was selected most often; however, choices 4 and 5 were also common. From a substantive point of view, it therefore seems that options in the middle, 3, 4 and even 5, are quite similar in the eyes of respondents. Thus, re-scoring seems reasonable.
	3	6	42	1.43	.62	
v24	1	3	2	.81	.18	Data codes 1 and 2 were merged because of very few answers in the case of option 1 and not too different ability mean scores. The new score value for these is 3 points, as data code 6, which is awarded 4 points, has a slightly higher ability mean, but not so much higher, however, that these codes should get only 2 points (compared to the difference in data codes 6 and 5 or 3). Experts' estimate was 4 and selections between 3 and 5 were therefore expected. Choice 4, however, was not most often chosen, which confirms that items between 3 and 5 seem quite similar to respondents. Thus, from a substantive point of view, re-scoring is justified.
	2	4=3	6	.94	.41	
	6	4	14	1.08	.39	
	5	5	26	1.21	.43	
	3	5	27	1.29	.49	
	4	6	25	1.56	.69	
v25	2	3	2	.72	.46	No changes were made. There is only a small number of respondents who selected option 2, but the ability mean for this was significantly different from data code 3 and, therefore, it was not reasonable to re-score this option. Experts' estimate was 5, and, from a substantive point of view, it did not need re-scoring even though option 6 was chosen almost as often as option 5.
	3	4	12	1.08	.44	
	6	5	13	1.14	.53	
	4	5	36	1.25	.47	
	5	6	37	1.43	.61	
v26	1	1	1	.29	.28	Data codes 3 and 4 were merged because of very few answers in the case of option 3 and similar ability means. The new score is 4; in addition, the score of data code 2 should increase by 1 because it has an ability mean significantly closer to the codes with a score of 4 rather than 1. Experts' estimate was 6 and it was selected most often.
	2	2=3	1	.81	.43	
	3	3=4	2	1.00	.11	
	4	4	11	.95	.42	
	5	5	17	1.20	.48	
	6	6	68	1.38	.55	

*all the response options selected by at least one participant for any item

**points given for each option if selected (if '=' then re-scored as a result of IRT analysis)

***how frequency the particular option has been chosen

****ability mean – the score that shows how difficult the particular option was to the respondents