Rasch Model of the COVID-19 Symptom Checklist - Towards a Higher Quality of Self-Reported Measurement

Tanja A Stamm (✉ Tanja.Stamm@meduniwien.ac.at)

Vienna Medical University <https://orcid.org/0000-0003-3073-7284>

Valentin Ritschl

Medical University of Vienna: Medizinische Universität Wien

Maisa Omara

Medical University of Vienna: Medizinische Universität Wien

Margaret R Andrews

Medical University of Vienna: Medizinische Universität Wien

Nils Mevenkamp

MCI: Management Center Innsbruck Internationale Hochschule GmbH

Angelika Rzepka

Austrian Institute of Technology GmbH

Michael Schirmer

Medical University of Innsbruck: Medizinische Universität Innsbruck

Siegfried Walch

MCI: Management Center Innsbruck Internationale Hochschule GmbH

Thomas Salzberger

Wirtschaftsuniversität Wien: Wirtschaftsuniversität Wien

Erika Mosor

Medical University of Vienna: Medizinische Universität Wien

Research article

Keywords: Coronavirus, COVID-19, SARS CoV-2, symptom checklist, self-reported instrument, fundamental principles of measurement

Posted Date: February 19th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-251823/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background: Inaccurate measurement of self-reported instruments including questionnaires and symptom checklists jeopardizes the comparability of the results. We therefore used advanced psychometric modelling to determine if the fundamental principles of measurement of an online self-reported COVID-19 symptom checklist were met or whether adaptations were necessary to increase measurement precision.

Methods: Fit to the Rasch model was examined in a sample of 1638 Austrian citizens who completed an online COVID-19 symptom checklist on up to 20 days during a period of restrictive country-wide COVID-19 measures.

Results: The longitudinal application of the self-reported COVID-19 symptom checklist increased the fit to the Rasch model. The items 'fatigue', 'headache' and 'sneezing' had the highest likelihood to be affirmed. The item 'cough' showed a significant misfit to the fundamental measurement model and an additional dependency to 'dry cough/no sputum production'. Several personal factors, such as gender, age group, educational status, COVID-19 test status, comorbidities, immunosuppressive medication, pregnancy and pollen allergy led to systematic differences in the patterns of how symptoms were affirmed. Adjustments ranged from ± 0.25 to ± 0.01 on the metric scales (0 to 10) to which the raw scores were transformed.

Conclusion: Except for some basic adaptations, the present analysis supports the combination of items. More accurate item wordings co-created with lay persons and adjustments for personal factors would increase measurement precision of the self-reported COVID-19 symptom checklist.

Introduction

The novel Coronavirus Disease 2019 (COVID-19) spread rapidly worldwide and the number of cases increased globally at an accelerated rate [1, 2]. While measures are needed to decrease the virus spreading and mitigate the impact of the pandemic, effective monitoring is essential to tailor these measures to the current situation. Self-reported symptom tracking contributes to monitoring [3] and a number of self-reported symptom checklists and trackers for COVID-19 exist; some of them also allow their users estimating the risk for a SARS-CoV-2 infection [4]. Moreover, self-reported symptom-tracking could increase patient safety at home in COVID-19 patients, who do not need hospitalisation. In oncology, even survival could be improved by collecting self-reported side-effects in a timely manner [5]. Self-reported data are thus a vital part of a comprehensive outcome measurement scheme and large international research initiatives are developing the technical, ethical and legal infrastructure to collect such data on country-level.

Self-reported patient data are acquired with specific instruments. To be fit-for-purpose, these tools must fulfill certain psychometric properties. Most commonly, psychometric properties are examined on the basis of the classical test theory [6–9]. However, the classical test theory focuses on overall, sample-based statistics, which provide little insight into how individual items actually work [9]. Furthermore, the classical test theory makes specific assumptions, such as normally distributed populations and interval-scaled response data, that are rarely met in practice. An alternative psychometric approach to assess whether an instrument fulfills the fundamental principles of measurement is the application of the Rasch model [10]. It overcomes some of the simplistic assumptions of the classical test theory and provides insight at the item level. Based on the Rasch model, items can be brought into a hierarchy according to their level of 'difficulty' or likelihood to be affirmed, which means that more 'difficult' items are only affirmed by persons with higher 'abilities'. A range of possible misfits to the Rasch model exists. A basic test compares actual responses to the expected responses based on parameter estimates. Another test checks, if the expected response to a specific item differs between respondents from different sub-groups of persons (e.g. females versus males, people from different age groups, or people with versus without co-morbidities), although having the same 'ability'/likelihood to affirm a certain number of items. This is referred to as differential item functioning [11]. If differential item functioning is apparent, parameter invariance is violated and findings between different groups of people cannot be compared without adjusting the scale. Related to parameter invariance is local dependency. Local dependency represents an additional dependency between items beyond the relationship related to the latent construct, which is measured by the instrument. Thus, local dependency also distorts the metric of the measure. Another fundamental requirement of measurement is unidimensionality [12] which applies whenever a scale is to be summarized by one measure.

Although a variety of self-reported COVID-19 symptom screening checklists exists [4], items are often not formulated in a standard manner and psychometric information on these instruments is scarce. This lack of knowledge on the psychometric properties of the symptom checklists prevents us from judging whether we use them in the right way or whether adaptations might be necessary. As more symptoms might increase the likelihood for a suspected case, this might be similar to the measurement of a latent construct. Furthermore, a recent Cochrane review concluded that individual COVID-19 signs and symptoms, such as cough, sore throat, fever, myalgia or arthralgia, fatigue, and headache, had very poor diagnostic power [13]. Moreover, inaccurate measurement in self-reported instruments jeopardizes the comparability of the results. We therefore used the Rasch model to determine if the fundamental principles of measurement of an online self-reported COVID-19 symptom checklist were met or whether adaptations were necessary to increase measurement precision.

Methods

We conducted a psychometric analysis using a sample of 1638 Austrian citizens who completed an online COVID-19 symptom checklist on up to 20 days during a period of restrictive country-wide COVID-19 measures. After the first confirmed cases of COVID-19 in Austria on 25 February 2020, nationwide infection control measures were ordered by the Austrian government from 16 March 2020 onwards. Public life in Austria remained severely affected before first easing measures were implemented in mid-April 2020 [14]. The self-reported, online COVID-19 symptom checklist used in the present study was therefore available from 22 March to 30 April, 2020.

We developed the checklist based on the WHO symptom descriptions for COVID-19 [15] and included fever, fatigue, cough, dry cough/no sputum production, pain in limbs, sore throat, headache, shortness of breath, chills, vomiting, diarrhea, nasal congestion, sneezing, sniffles/rhinitis and smell and taste disorders (Supplement Table A). Response options and scores for each item were as follows: 'Yes' (scoring 1), 'No' (0) and 'I can't say that' (3). For the analysis, we dichotomized the items by collapsing 'No' and 'I can't say that' because we assumed that these two answers would indicate that the participant had not experienced a certain symptom. In addition, we asked the participants to state gender, age group, the highest completed education, the current smoking status, body height and weight, whether any type of COVID-19 test had been conducted and if so, what the result of this test was, whether any comorbidities existed (nervous system, cardiovascular, gastrointestinal, liver, kidney, oncologic, high blood pressure and/or diabetes), whether he or she was taking immunosuppressive medication and whether the participant was pregnant (females only). We used a self-reported code consisting of numbers and string based on first name initials, initials of first names of relatives and birth months to allocate multiple assessments to the correct participants despite guaranteeing anonymity. Due to the psychometric nature of this study, only complete cases were included. The relevant ethical committees approved the study (Medical University of Vienna 1379/2020, Medical University of Innsbruck 1076/2020 and ethical committee of the region Vorarlberg).

Fit to the Rasch measurement model

Overall and item-based fit to the Rasch model was explored in a series of dichotomous models [16] using two different data sets: one data set with the questionnaire filled in for the first time when the participants entered the study and another dataset in which we recorded a symptom to be affirmed by a participant, if it was ticked at least once during the period of nationwide restrictive measures. We used raw scores without weighting [17] and determined the hierarchy of items based on their location parameters. The item local parameter refers to the likelihood of each item to be affirmed. Items differ in their likelihood to be affirmed. A hierarchy of items can be determined based on this likelihood. Likewise, a hierarchy of persons can be established based on the likelihood that a person is likely to affirm more or less items.

Item fit residuals between -2.5 and $+2.5$ with non-significant F-tests represented individual item fit. Non-significant chi-squared values were interpreted as fit to the latent trait. Local dependency between items was determined using residual correlations based on a cut-off of 0.2 above the mean [18]. To assess the instrument's item-based internal consistency and reliability, we compared Cronbach's alpha with the person separation index (PSI). The PSI refers to the reproducibility of relative measure location and indicates whether a scale is able to distinguish between people with higher and lower levels of the concept

measured by the instrument [19]; in general, a $PSI \geq 0.7$ indicates that the instrument is sufficiently suitable for group comparisons.

Unidimensionality

To test unidimensionality, we used an approach proposed by Smith [20] and combined principal component analysis of the item residuals with a series of t-tests to assess whether subsets of residuals which loaded positively or negatively resulted in different estimates of person parameters. These sets of items were chosen as a way to maximize the contrast between them and were thus then most likely to violate the assumption of unidimensionality [21].

Differential item functioning

Differential item functioning was assessed separately for each item by comparing the expected responses to a specific item between respondents from different sub-groups who shared the same likelihood to affirm a certain number of items. The sub-groups were built based on gender (female, male, divers/other), age group (10 sub-groups listed in Table 1), highest completed education (6 levels listed in Table 1), COVID-19 test status (pos/neg/no test), comorbidities (yes/no), immunosuppressive medication (yes/no), pregnancy (yes/no), current smoking status (yes/earlier, but not now/never) and body mass index (BMI; above versus below median). As only seven participants indicated 'divers/other' for gender, we did not include this category into the differential item functioning analysis and compared only female and male participants. If differential item functioning was apparent in an item for a personal factor with more than two properties, we determined between which sub-groups these differences occurred using post hoc analysis of the residual means.

Table 1
Participant characteristics

Personal factors/characteristics	Frequencies
Total number of participants N (%)	1638 (100%)
Gender n (%)	
Female	1088 (66.4%)
Male	543 (33.2%)
Divers/other	7 (0.4%)
Pregnancy (women only) n (%)	17 (1.6% of the women)
Age groups n (%)	
0–9 Years	21 (1.3%)
10–19 Years	31 (1.9%)
20–29 Years	421 (25.7%)
30–39 Years	451 (27.5%)
40–49 Years	262 (16.0%)
50–59 Years	264 (16.1%)
60–69 Years	137 (8.4%)
70–79 Years	40 (2.4%)
80–89 Years	10 (0.6%)
≥ 90 Years	1 (0.1%)
Highest education n (%)	
Unfinished compulsory education	37 (2.3%)
Completed compulsory education	30 (1.8%)
Completed apprenticeship	140 (8.5%)
Completed post-secondary non-tertiary education	453 (27.7%)
Completed first stage of tertiary education	718 (43.8%)
Completed second stage of tertiary education	260 (15.9%)
COVID-19 tested n (%)	
Positive	16 (1%)
Negative	187 (11.4%)
Not tested	1435 (87.6%)
Comorbidities n (%)	
Yes	359 (21.9%)
No	1279 (78.1%)
<i>Note. N (%) = total number of participants (percentage); n (%) number of participants (percentage)</i>	

Personal factors/characteristics	Frequencies
Immunosuppressive medication n (%)	
Yes	45 (2.7%)
No	1593 (97.3%)
<i>Note. N (%) = total number of participants (percentage); n (%) number of participants (percentage)</i>	

Person-item targeting

Person-item targeting was inspected graphically using person-item map.

Transformation to a metric interval scale

Based on the logit scale from the Rasch model, we transformed the raw sum scores into a metric scale. If differential item functioning existed for a personal factor, we split the respective item and performed a separate metric transformation for each sub-group. We used the differences between these metric scales to adjust the scores for the respective personal factor, e.g. for people with and without comorbidities. All analyses were performed with either Microsoft Excel, RUMM2030 or the eRm and ltm packages in R (www.r-project.org).

Results

Participants from all age groups, ranging from 0 to 9 years to ≥ 90 years, filled in the symptom checklist. Two thirds of the participants (66%) were female (Table 1). Twelve participants (0.7%) indicated fever at least once during the period of nationwide restrictive measures (8 [0.5%] when they first filled in the questionnaire); 945 (57.7%; first data entry: 459; 28%) reported fatigue, 409 (25%; first data entry: 158; 9.6%) cough, 90 (5.5%; first data entry: 76; 4.6%) dry cough/no sputum production, 238 (14.5%; first data entry: 91; 5.6%) pain in limbs, 482 (29.4%; first data entry: 116; 7.1%) sore throat, 786 (48%; first data entry: 203; 12.4%) headache, 181 (11.1%; first data entry: 70; 4.3%) shortness of breath, 98 (6%; first data entry: 14; 0.9%) chills, 133 (8.1%; first data entry: 21; 1.6%) vomiting, 308 (18.8%; first data entry: 74; 4.5%) diarrhea, 604 (36.9%; first data entry: 317; 19.4%) nasal congestion, 752 (45.9%; first data entry: 361; 22%) sneezing, 436 (26.6%; first data entry: 229; 14%) sniffles/rhinitis and 108 (6.6%; first data entry: 81; 4.9%) smell and taste disorders.

Diagnosis of measurement problems

The data set with the first responses of the participants (model 1 in Table 2) showed a considerable misfit to the Rasch model and a substantial discrepancy between PSI (-0.06) and Cronbach's alpha (0.68). The second dataset, where a symptom was recorded as affirmed, if it was scored at least once during the observation period (model 2 in Table 2), showed a better model fit (a lower Root Mean Square Error of Approximation) and a substantially smaller difference between PSI (0.6) and Cronbach's alpha (0.75). Due to the better model fit and the fact that fit statistics of symptom checklists could be affected if only few persons of a population report symptoms, we decided to use only the second data set for further analyses, where a symptom was recorded as affirmed, if it was scored at least once during the observation period. In addition, item fit statistics on the first responses of the participants are shown in Supplement Table B.

Table 2

Model fit statistics. Mean item log residual test of fit, item-trait interaction chi-square statistics and Root Mean Square Error of Approximation (RMSEA) were calculated to assess model fit. **Model 1** ("First") refers to the questionnaires filled in for the first time when the participants entered the study. In **Model 2** ("Ever"), a symptom was considered affirmed by a participant, if it was ticked at least once during the period of nationwide restrictive measures

Model fit statistics							Unidimensionality analysis				
	Mean item location ($\pm$ SD)	Mean item fit residual ($\pm$ SD)	Mean person location ($\pm$ SD)	Mean person fit residual ($\pm$ SD)	Person separa- tion index (PSI)	Cron- bach's α	Root Mean Square Error of Approx- imation (RMSEA)	Number of sign. t-tests	Sample	% of sign. t- tests	Lower bound of 95% CI
Model 1. "First"	0 ($\pm$ 1.42)	-1.19 ($\pm$ 2.33)	-2.89 ($\pm$ 1.04)	-0.3 ($\pm$ 0.52)	-0.06	0.68	0.052	14	1638	0.9%	-0.2%
Model 2. "Ever"	0 ($\pm$ 1.59)	-0.86 ($\pm$ 1.49)	-1.91 ($\pm$ 1.37)	-0.26 ($\pm$ 0.57)	0.60	0.75	0.027	43	1638	2.6%	1.6%
<i>CI = Confidence interval</i>											

Only one item (cough) had significantly deviating F-tests with an item-based fit residual being below - 2.5 and a significantly deviating chi-squared value (Table 3). Fatigue, headache and sneezing were the items with the highest likelihood to be affirmed, compared to fever with the lowest probability of affirmation by the participants, followed by dry cough and chills.

Table 3

Item fit statistics sorted in an ascending order according to item location in the data set with the items affirmed, if they were scored positively at least once by each participant during the observation period. The smallest (negative) item location of fatigue implies that it was the item with the highest likelihood to be affirmed, whereas fever showed the lowest likelihood. Only one item (item 3 cough) showed a fit residual below the threshold of -2.5 and a significant F-test which represented individual item misfit. Cough also had a significant chi-squared value which can be interpreted as misfit to the latent trait. The Bonferroni corrected significance level was 0.000667. Significances are highlighted in bold letters. Differential item functioning (DIF; depicted in the far right column) exists, if the expected response to a specific item differs between respondents from different sub-groups of persons (e.g. females versus males, people from different age groups or people with and without co-morbidities), even though they have the same 'ability'/likelihood to affirm the same number of items.

Item number	Description	Location	Standard error	Fit residual	F-stat	p-value	ChiSq	p-value	DIF for
2	Fatigue	-2.47	0.06	-0.47	0.41	0.9284	4.45	0.7271	Age; educational status
7	Headache	-1.87	0.06	0.11	1.70	0.0839	14.29	0.0463	Gender; comorbidities
13	Sneezing	-1.75	0.06	1.28	2.02	0.0342	17.95	0.0122	Pollen allergy
12	Nasal congestion	-1.26	0.06	-1.96	2.63	0.0051	20.61	0.0044	Pollen allergy
6	Soare throat	-0.84	0.06	0.11	0.68	0.7264	7.23	0.4056	-
14	Sniffles/rhinitis	-0.62	0.07	-2.59	2.96	0.0017	23.54	0.0014	Gender
3	Cough	-0.60	0.07	-3.55	4.73	0.0000	38.43	0.0000	-
11	Diarrhea	-0.15	0.07	1.61	1.38	0.1928	15.32	0.0322	-
5	Pain in limbs	0.24	0.08	-1.55	1.78	0.0667	17.71	0.0134	Age
8	Shortness of breath	0.65	0.08	-2.37	1.24	0.2688	13.35	0.0641	Comorbidities
10	Vomiting	1.01	0.09	-1.41	1.17	0.3094	12.41	0.0879	Pregnancy
15	Smell and taste disorders	1.21	0.10	0.47	1.33	0.2188	14.21	0.0475	COVID-19 test
9	Chills	1.44	0.11	-1.73	0.92	0.5077	9.67	0.2081	-
4	Dry cough/no sputum production	1.47	0.11	-1.10	1.16	0.3164	11.26	0.1275	Age
1	Fever	3.55	0.25	0.24	0.98	0.4555	10.44	0.1652	

Adjustment for differential item functioning

All personal factors, except for smoking status and BMI, led to differential item functioning in certain items (Table 3). As expected, taste and smell disorders were more likely to be affirmed by participants with a positive COVID-19 test result, participants with pollen allergy were more likely to indicate nasal congestion and sneezing and pregnant women were more likely to experience vomiting. However, less obvious was that participants with comorbidities exhibited a higher likelihood to indicate shortness of breath and were less likely to score headache when compared to participants without comorbidities; pain in the limbs was more often affirmed by persons with immunosuppressive medication than persons without; women were more likely to indicate headache when compared to men, and men indicated more often rhinitis/sniffles than women (Fig. 1). Participants with higher educational status (completed post-secondary non-tertiary education or first stage of tertiary education) were more likely to indicate fatigue than people with completed apprenticeships (the significant absolute residual mean differences in the post hoc analysis were 0.29 and 0.4, respectively) and unfinished compulsory education (the significant absolute residual mean differences were 0.5 and 0.61, respectively). Likewise, people with completed second stage of tertiary education were more likely to affirm fatigue in comparison to participants with unfinished compulsory education (the significant absolute residual mean difference was 0.5). Fatigue was also most prevalent in younger adults (from 20 to 49 years of age), when compared to both children/adolescents up to 19 and older adults of 50 + years (Supplement Figure A). However, as the highest completed

education depended also on the age group, e.g. participants below the age of 20 could not have completed the second stage of tertiary education, and the numbers of participants in the age groups were heterogeneous (ranging from 1 to 541; Table 1), we did not further assess differences between specific age groups.

Local dependency in relation to item fit

Local dependency was detected between items 3 (cough) and 4 (dry cough/no sputum production), items 12 (nasal congestion) and 13 (sneezing), as well as items 12 (nasal congestion) and 14 (sniffles/rhinitis). The additional 'local' dependency of nasal congestion, sneezing and/or sniffles/rhinitis seemed to be rather evident and well known. Moreover, these three symptoms might be less important regarding COVID-19 than, for example, cough. Cough and dry cough also showed such an additional dependency which might also relate to the misfit of cough (item 3) in Table 3. Finally, we found both data sets to be unidimensional (the last four far right columns in Table 2 show these results).

Person item targeting

From the graphical inspection of the person-item map (Supplement Figure B), it is evident that the items in general cover the range of symptoms in the population. We used the dataset where an item was recorded as affirmed, if indicated at least once during the observation period, which consisted of less zero-scored symptoms compared to the data set of the questionnaires filled in only once.

Transformation to a metric interval scale

The raw sum scores and the corresponding values on an interval metric scale from zero to ten are depicted in Table 4. Higher scores indicate a more symptoms. We split and adjusted items 'headache', 'shortness of breath' and 'rhinitis/sniffles', 'pain in limbs' and 'fatigue' according for the personal factors comorbidities, gender, immunosuppressive medication and educational status (Table 4; details of the calculations are depicted in Supplement Table C). Adjustments for personal factors differed in magnitude over the range of each metric scale.

Table 4

Transformations of raw scores into metric scales (logit scales and from zero to ten). 'COM' refers to persons with comorbidities; 'IMM' refers to persons with immunosuppressive medication; The numbers below COM, IMM and gender represent the difference between the metric scales for these subgroups, e.g. persons with and without comorbidities, and represent a possibility to adjust the metric scores for a better comparability of persons with different personal factors.

All - no split		COM			IMM		Gender		Education level	
Raw score	Metric scale	Zero to ten	With	Without	With	Without	Female	Male	Lower than post secondary	≥ Post secondary
0	-4.02	0.55	0.06	-0.06	-0.08	0.08	-0.08	0.08	0.25	-0.25
1	-3.10	1.56	0.06	-0.06	-0.09	0.09	-0.07	0.07	0.19	-0.19
2	-2.40	2.34	0.04	-0.04	-0.11	0.11	-0.05	0.05	0.15	-0.15
3	-1.87	2.93	0.02	-0.02	-0.13	0.13	-0.02	0.02	0.11	-0.11
4	-1.42	3.43	0	0	-0.14	0.14	0	0	0.08	-0.08
5	-1.01	3.88	-0.03	0.03	-0.16	0.16	0.02	-0.02	0.06	-0.06
6	-0.62	4.31	-0.06	0.06	-0.16	0.16	0.03	-0.03	0.05	-0.05
7	-0.24	4.73	-0.09	0.09	-0.16	0.16	0.04	-0.04	0.03	-0.03
8	0.13	5.15	-0.11	0.11	-0.15	0.15	0.03	-0.03	0.03	-0.03
9	0.51	5.57	-0.12	0.12	-0.14	0.14	0.03	-0.03	0.02	-0.02
10	0.91	6.01	-0.13	0.13	-0.12	0.12	0.03	-0.03	0.01	-0.01
11	1.34	6.48	-0.12	0.12	-0.09	0.09	0.02	-0.02	0.01	-0.01
12	1.82	7.02	-0.11	0.11	-0.07	0.07	0.02	-0.02	0.01	-0.01
13	2.42	7.68	-0.09	0.09	-0.05	0.05	0.01	-0.01	0	0
14	3.28	8.63	-0.06	0.06	-0.03	0.03	0.01	-0.01	0	0
15	4.52	10	0	0	0	0	0	0	0	0

Discussion

For the first time, we assessed measurement precision of a COVID-19 symptom checklist using advanced psychometric methods and suggested some basic adaptations which could increase its precision. Local dependency between the items 'cough' and 'dry cough/no sputum production', for example, showed the need for more accurate item wordings which would make it clearer to the participants how to differentiate between these items. The unclear difference between what is meant with 'cough' and 'dry cough/no sputum production' could have prevented them from precise scoring and an improved item wording would be preferable. Likewise, internationally agreed standard item wordings for COVID-19 symptom checklists co-created with lay persons could improve comparability of the findings, especially across datasets and different countries. Such an agreed version should then also be tested for measurement precision using an approach similar to ours.

Our analysis showed that participants with certain characteristics, such as comorbidities, intake of immunosuppressive medications, gender and educational status, scored systematically differently on some items, although having in principle the same likelihood to affirm the respective items. This means for example that people with comorbidities had a higher likelihood to affirm shortness of breath due to this confounder. Likewise, people with higher educational status were more likely to score fatigue despite their otherwise similar likelihood to affirm a similar number of items when compared to people with a lower educational status. The symptoms on the COVID-19 checklists could also occur due to other diseases or in relation to certain

personal characteristics. An adjustment of scores according to these characteristics would increase the comparability of the findings between individuals from different sub-groups.

The low PSI in the dataset with the questionnaires filled in for the first time by the participants could be related to the small number of people who affirmed the items when they first responded to the questionnaire. It might thus be difficult to determine measurement precision of symptom checklists, if only a small number of people report symptoms. We recorded similar findings in previous studies where the large number of zero-scored symptoms led to right skewed distributions and a unfavorable person-item targeting [21, 22]. In general, measurement precision of self-reported symptom checklists, especially in case of easily fluctuating symptoms, such as the ones used in our COVID-19 checklist, might be increased by using these checklists longitudinally.

We could show that fever, dry cough and chills had a low probability of affirmation. This could be related to the fact that participants were anxious to report these symptoms in a digital way, even if anonymity was guaranteed and the participation in our study was voluntary. Participants could be reluctant to fill in such symptoms in digital tools and might fear that this could lead to personal consequences, such as quarantine or other restrictions. This fact needs to be taken into account when self-reported checklists are used and could be a reason for the very poor diagnostic power [13].

As we used Austrian data in our study, differential items functioning between countries and the assessment of cross-cultural differences could thus not be assessed. In terms of future work, it would be interesting to repeat the Rasch model in an international data set collected by an online self-reported COVID-19 symptom checklist with adapted wording of specific items.

Conclusion

Apart from some basic adjustments, the analysis supports the present combination of items into a comprehensive COVID-19 symptom questionnaire. More accurate item wordings co-created with lay persons and adjustments for personal factors would increase measurement precision of the self-reported COVID-19 symptom checklist.

Declarations

Funding Source: This study was partly funded by the COVID-19 Rapid Response Funding Scheme of the Vienna Science and Technology Fund (project number COV20-028). The funding institution had no influence on the results of this work.

Ethics approval and consent to participate

The relevant ethical committees approved the study (Medical University of Vienna 1379/2020, Medical University of Innsbruck 1076/2020 and ethical committee of the region Vorarlberg). All participants consented in written before filling in the questionnaire.

Consent for publication

Not applicable

Competing interests

We declare no competing interests related to this work.

Funding

This study was partly funded by the COVID-19 Rapid Response Funding Scheme of the Vienna Science and Technology Fund (project number COV20-028). The funding institution had no influence on the results of this work.

Authors' contributions

VR, MO, RA, NM, AR, MS, SW, ToS, EM and TS were responsible for the study conceptualization and design. VR, EM and TS designed the online survey and analysed the data. VR, MO, RA, NM, AR, MS, SW, ToS, EM and TS interpreted the data, wrote and reviewed the manuscript. EM and TS were responsible for the visualization, including tables and figures.

Acknowledgements

We would like to thank all the participants of the study.

Availability of data and material (data transparency)

Data are available from the corresponding author upon reasonable request.

References

1. Zhu N, et al., *A novel coronavirus from patients with pneumonia in China, 2019*. New England Journal of Medicine, 2020.
2. World Health Organization. *Clinical management of severe acute respiratory infection when novel coronavirus (2019-nCoV) infection is suspected—interim guidance*. 2020 [cited 2020 05.04.2020]; Available from: [https://www.who.int/publications-detail/clinical-management-of-severe-acute-respiratory-infection-when-novel-coronavirus-\(ncov\)-infection-is-suspected](https://www.who.int/publications-detail/clinical-management-of-severe-acute-respiratory-infection-when-novel-coronavirus-(ncov)-infection-is-suspected).
3. Menni C, et al., *Real-time tracking of self-reported symptoms to predict potential COVID-19*. Nature medicine, 2020: p. 1–4.
4. Boulos MNK, Geraghty EM, *Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: how 21st century GIS technologies are supporting the global fight against outbreaks and epidemics*. 2020, BioMed Central.
5. Basch E, et al. Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial. J Clin Oncol. 2016;34(6):557.
6. Jokovic A, Locker D, Guyatt G, *Short forms of the Child Perceptions Questionnaire for 11–14-year-old children (CPQ 11–14): development and initial evaluation*. Health and quality of life outcomes, 2006. 4(1): p. 4.
7. Jokovic A, et al. Validity and reliability of a questionnaire for measuring child oral-health-related quality of life. Journal of dental research. 2002;81(7):459–63.
8. Jokovic A, et al. Questionnaire for measuring oral health-related quality of life in eight-to ten-year-old children. Pediatric dentistry. 2004;26(6):512–8.
9. Thomson WM, et al. Psychometric assessment of the short-form Child Perceptions Questionnaire: an international collaborative study. Community Dent Oral Epidemiol. 2016;44(6):549–56.
10. Rasch G, *On General Laws and Meaning of Measurement in Psychology*. Proceedings of the Fourth Berkley Symposium on Mathematical Statistics and Probability, 1961. 4: p. 321–333.
11. Tutz G, Schauberger G. A penalty approach to differential item functioning in Rasch models. Psychometrika. 2015;80(1):21–43.
12. Heene M, Kyngdon A, Sckopke P. Detecting Violations of Unidimensionality by Order-Restricted Inference Methods. Frontiers in Applied Mathematics Statistics. 2016;2:3.
13. Struyf T, et al., *Signs and symptoms to determine if a patient presenting in primary care or hospital outpatient settings has COVID-19 disease*. Cochrane Database of Systematic Reviews, 2020(7).
14. Pollak M, Kowarz N, Partheymüller J. *Chronology of the Corona Crisis in Austria - Part 1: Background, the way to the lockdown, the acute phase and economic consequences*. 2020 [cited 2020 30.8.]; Available from: <https://viecer.univie.ac.at/en/projects-and-cooperations/austrian-corona-panel-project/corona-blog/corona-blog-beitraege/blog51/>.
15. Organization WH. Report of the WHO-China joint mission on COVID-19, 16–24 February 2020. Geneva (Switzerland): World Health Organization; 2020.
16. Tennant A, Pallant J, *Unidimensionality matters!(A Tale of Two Smiths?)*. Rasch Measurement Transactions, 20 (1), 1048–1051. 2006.

17. Caplan DJ, Slade GD, Gansky SA. Complex sampling: implications for data analysis. J Public Health Dent. 1999;59(1):52–9.
18. Andrich D, *Rasch models for measurement*. Vol. 68. 1988: Sage.
19. Tennant A, Conaghan PG, *The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper?* Arthritis Care & Research, 2007. **57**(8): p. 1358–1362.
20. Smith EV Jr, *Understanding Rasch measurement: Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals*. Journal of applied measurement, 2002.
21. Stamm TA, et al. Rasch model of the child perceptions questionnaire in multi-country data. J Dent. 2020;93:103267.
22. Omara M, et al., *Rasch model of the Child Perceptions Questionnaire for oral health-related quality of life: A step forward toward accurate outcome measures*. J Am Dent Assoc, 2019.

Figures

Figure 1

Differential item functioning for comorbidities, immunosuppressive medication and gender. The bar chart charts on the left depict the absolute and relative frequencies of answers regarding the two items in persons with and without comorbidities and immunosuppressive medication, as well as gender. While persons with comorbidities were more likely to indicate shortness of breath (24%) compared to people without comorbidities (1%), they were less likely to affirm headache (45% without compared to 49% with comorbidities). Persons with immunosuppressive medication were more likely to indicate pain in limbs (47%) compared to people without (14%). Women were more likely to indicate headache (55% versus 34% in men) and men rhinitis/sniffles (28% versus 26% in women). The item characteristic curves in the right column were split for people with comorbidities/immunosuppressive medication (red curves) and without (blue curves), as well as women (blue curves) and men (red curves).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementTablesABFiguresAB.pdf](#)
- [SupplementTableC.pdf](#)