Heliyon 8 (2022) e10311

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

# Satisfaction towards virtual courses: Development and validation of a short measure in COVID-19 times

José Ventura-León *, Tomás Caycho-Rodríguez, Jency Mamani-Poma, Lucerito Rodriguez-Dominguez, Luciana Cabrera-Toledo

*Universidad Privada del Norte, Peru*

ARTICLE INFO

*Keywords:*
Satisfaction
Virtual courses
Validation
Item response theory
University students

ABSTRACT

This study was aimed to develop and validate a short scale to measure satisfaction with virtual courses (SVC–S) in a sample of higher education students during the covid-19 pandemic; specifically, in the year 2021. A total of 3080 students between 16 and 56 years of age participated (Mean = 25.71; *SD* = 8.83); 1836 were female (59.60 %) and 1244 male (40.40 %). The participants were students from three cities in Peru (77.90% from Lima, 12.70% from Trujillo and 9.42% from Cajamarca). Qualitative and quantitative procedures were followed for the construction of the SVC-S. Item response theory (IRT) considering Samejima's two-parameter Graded Response Model (GRM) (2PL) and the test-item information function was used to establish accuracy/reliability, and the relationship of the SVC-S with a similar measure was examined to demonstrate convergence and discrimination. The results reveal that the data present an optimal fit (M2 (2) = 3.62; RMSEA = .016; CFI = 1.00). Reliability is excellent ($r_{xx}$ = .93) and the information function suggests that the instrument is more accurate at low levels of the latent trait. Regarding convergence with an academic satisfaction scale, the SVC-S showed an appropriate correlation ($r$ = .70) whose average variance extracted (AVE) reported good discrimination of the constructs; despite being conceptually similar. SVC-S is concluded to be a valid and reliable measure that can be used in future studies in higher education.

## 1. Introduction

The COVID-19 pandemic is a health emergency that had strong repercussions on the social, economic, and educational aspects of society. As for education, the global spread of the disease led to the suspension of on-face classes from the beginning of the pandemic. Previously, infectious diseases such as SARS and H1N1 led to the suspension of traditional educational activities, adversely affecting educational activities in many countries (Cauchemez et al., 2014). The suspension of on-face classes by COVID-19 affected about 1.5 million students in 2020 in mid-April 2020, a figure that decreased to 900 million affected students in June 2020; while, at the beginning of 2021, 250 million students were still affected by the closure of schools and universities (Tang et al., 2021). Thus, the universities had to face an abrupt change from face-to-face to remote (Tejedor et al., 2020), which became the only way to continue university classes and minimize the impact of the pandemic to the academic development of students (Hassan et al., 2021). Peru was one of the many countries that decided to close schools and universities to try to contain the spread of the virus, taking into consideration that social distancing has been the most effective preventive strategy against COVID-19 until the arrival of the vaccine (del Rio and Malani, 2020).

In this scenario, the suspension of in-person activities caused a strong pressure on teachers and students to adapt to virtuality, redesigning the subjects and supporting the demand for online courses (González-Calvo et al., 2020). In this way, educational institutions adopted different teaching techniques, such as online lectures, audio and video recorded lectures, and online shared materials, among others (Favale et al., 2020). They also used different online evaluation methods, such as questionnaires, exams, and online activities (George, 2020). Unfortunately, a large number of educational institutions, teachers, and students were not adequately prepared for this new experience. In Peru, for example, at the beginning of the pandemic, about 70% of universities had no previous experience in conducting virtual courses, and only some used electronic platforms and digital resources, but not all of them (Figallo et al., 2020). Both teachers and students faced a wide range of logistical, technical, financial, and social problems (Chakraborty et al., 2021). For example, a

* Corresponding author.
*E-mail address:* jose.ventura@upn.pe (J. Ventura-León).

https://doi.org/10.1016/j.heliyon.2022.e10311
Received 10 February 2022; Received in revised form 5 July 2022; Accepted 11 August 2022

large number of students do not have access to the Internet, which could worsen equity problems in the long run. In the case of teachers, there has been great concern about preparing the necessary infrastructure for online classes, access to the Internet and having adequate space (Tang et al., 2021). The shift to online learning and the limitations noted have affected students, educators, and learning performance in general (Ustun, 2021).

After almost two years, the COVID-19 pandemic is still not under control in many countries, leading to an increased risk of COVID-19 becoming endemic (Hall et al., 2020). This makes it very likely that remote teaching and learning will continue for several more months or years (Aguilera-Hermida et al., 2021). In this sense, the debate about the quality of learning and student satisfaction with online learning has been initiated. Therefore, it is important to understand how satisfied students are with online courses. This would provide higher education institutions with valuable information to establish properly planned instructional methods during this pandemic and to prepare for future emergencies, such as other pandemics or natural disasters (Baber, 2020). The importance of the academic satisfaction of students lies in their relationship to academic performance, retention of information, and continuous effort during learning (Zeng and Wang, 2021). However, recent studies show that e-learning during the COVID-19 pandemic led to a decrease in students' academic satisfaction (Gonçalves et al., 2020; Hamdan et al., 2021). Variations in students' satisfaction with online classes are directly and indirectly affected by their self-efficacy in using computer tools, perceived ease of use, and usefulness of online teaching platforms (Jiang et al., 2021). Similarly, class time, loss of interest, motivation, and use of online exams as a means of assessment are also factors that significantly affect student satisfaction with online classes (Basuony et al., 2021). In addition, it has previously been noted that more than half of the students enrolled in online programs drop out due to dissatisfaction with the quality of instruction (Betts, 2009). On the other hand, the presence of students satisfied with online education has also been reported, considering it relevant to their learning needs and a way to cope with COVID-19 anxiety (Agarwal and Kaushik, 2020). Unfortunately, studies often do not provide sufficient information on how the level of satisfaction was measured (Almusharraf and Khahro, 2020).

Regarding the measurement of satisfaction with online classes during the pandemic, *ad hoc* questions have often been developed that have not demonstrated evidence of validity and reliability (e.g, Chen et al., 2020), while, at other times, long measures have been used to assess the level of student satisfaction with specific aspects of teaching, such as the effectiveness of the teacher's work (Fatani, 2020). Likewise, other studies mention the adaptation of scales and indicate that psychometric studies were conducted, but do not provide evidence of this (for example, Jiang et al., 2021). Given this methodological gap, and the need to assess student satisfaction with online learning during the pandemic, this study aimed to develop and validate a short measure of satisfaction with online courses (SVC–S), in order to provide a measurement instrument that allows exploration, identification, and relationship with other academic variables. The SVC-S was developed on the basis of the theoretical models of Oliver (2010), Winston and Sommers (2013), Hill et al. (2007), Rai (2012), Ragin (2017). All these models understand educational institutions as an organization, where students are seen as recipients of a service. In this sense, they conceptualize satisfaction as the consumer's perception that their needs and desires have been met. A summary of the theoretical models can be seen in Table 1; additionally, this phase is used for item development and the results achieved are described in more detail there.

Unlike traditional psychometric studies, the SVC-S was developed based on the Item Response Theory (IRT). IRT is increasingly used in the construction of scales to measure educational and psychological attributes (Edelen and Reeve, 2007), as they allow more accurate assessments of the characteristics of each item compared to those reported by Classical Test Theory (CTT, Hambleton, 1989, van der Linden and Hambleton, 1997). IRT models complement for the evaluation of the dimensionality of a scale and the accuracy of the measurement at the item level (DeMars, 2010). The application of IRT allows obtaining information that CTT models cannot report, such as item parameters and reliability estimation along the continuum representing the measured latent trait. Regarding the parameters, the two-parameter logistic model (2PL), assumes the presence of a single underlying trait and difficulty (β) and discrimination (α) parameters for the items. The difficulty of an item describes the amount of the latent trait that the item requires to be answered; whereas discrimination allows us to obtain information about which items allow us to differentiate between people with different levels of the trait. Also, the difficulty and discrimination parameters can be estimated independently of the trait level of the examinee. This is a strength that CTT-based models do not have, where the statistical parameters of the items depending on the characteristics of each sample. On the other hand, IRT models do not assume that items are equally informative across the range of latent traits, but rather that one item may provide more or less information than another item (Embretson and Reise, 2000).

**Table 1.** Comparative table of different definitions of satisfaction.

| FAMILIARIZATION AND SEGMENTATION | | | | |
|---|---|---|---|---|
| Oliver (2010, p.8) | Winston and Sommers (2013) | Hill et al. (2007, p. 31) | Rai (2012, p.105) | Ragin (2017, p. 379). |
| The satisfaction is the *response of fulfilment of consumer.* This is a judgment that a feature of the product/service, or the product or service itself, provides (or is providing) *a pleasant level of satisfaction related to consumption,* including levels of underuse or overuse. | *The satisfaction of consumer is the perception of consumer*, in and out of the organization, *of which his/her needs and wishes have been met in relation to his/her health care, and that he/she feels that he/she has received a treatment and high-quality services ranged as excellent.* | *If the product is fitted to expectations, the consumer is met*; if it is overcome, the consumer is very satisfied; if it is insufficient, the consumer is unsatisfied | The satisfaction of customer or *consumer is an emotional or cognitive response of the buyer* after the *assessment and comparison of the expectations prior to the purchase and real performance after consuming* the product or service during the transaction with an organization. | The satisfaction of the consumer (patient) *as a comparison between subjective assessment of a person from his/her expectations as for attention* and *perceptions of this person from the received real attention real.* The received *real attention is measured by the emotional and cognitive interaction of the consumer with the provider.* |
| CATEGORIZATION | | | | |
| Satisfaction: Subjective assessment involving a pleasant response that is got when the person has positive expectations. | | | | |
| CORRESPONDENCE | | | | |
| 1. The virtual courses are interesting (pleasant assessment).<br>2. The virtual courses meet my expectations (subjective assessment of the expectations).<br>3. The virtual courses are nice for me (pleasant assessment).<br>4. The virtual courses are a good alternative for learning (pleasant assessment).<br>5. In general, I feel satisfied with the virtual courses (global assessment). | | | | |

Note: Italics denote the relevant segment (segmentation). The different types of underlining indicate the similarity of the information segments (Categorization).

In addition, IRT allows assessment of the measurement accuracy of a test through the test information function (TIF), which assesses test accuracy at different levels of trait measurement, rather than providing a single value for reliability, such as that obtained through coefficient alpha (Embretson and Reise, 2000; Hambleton et al., 1991). In this sense, the more information the test provides at a specific ability level, the lower the error associated with the trait estimate and the higher the reliability of the test. IRT models are extremely useful for assessing metric properties of short instruments, provided that an adequate fit between the model and the data is observed (Embretson and Reise, 2000). In conclusion, having a measure of student satisfaction with virtual courses with adequate psychometric evidence will allow educators and educational psychologists to develop appropriate online learning programs that provide students with better learning environments.

## 2. Method

### 2.1. Participants

Initially, 3169 higher education students participated from a single Peruvian university during the year 2021. However, after the review of response patterns using the index *Zh* (Drasgow et al., 1985) because they presented *Zh* ± 2.0 (Felt et al., 2017) and these can be seen as outliers that may interfere with the estimation of the factor model (Yuan and Zhong, 2013). Thus, some cases were withdrawn, leaving a total of 3080 participants. Ages ranged from 16 to 56 years (Mean = 25.71; *SD* = 8.83); 1836 were female (59.60%) and 1244 male (40.40%). The participants were students from three cities in Peru (77.90% from Lima, 12.70% from Trujillo and 9.42% from Cajamarca) from the undergraduate level. The inclusion criteria included all students enrolled during the 2021 semester at a university in the northern sector of Lima, from the first to the last cycle, who agreed to participate in the study voluntarily. On the other hand, students with unusual patterns were excluded through the *Zh* index. The sample size was estimated using the 'semPower' library (Moshagen and Erdfelder, 2016) establishing *a priori* 5 degrees of freedom [k (k-3)/2]; RMSEA = .05; power of .95 and an alpha of .05, giving a total of 2062 observations; thus, the study exceeded the estimated sample value. The sampling was purposive because a set of participants was purposively chosen to respond to the study objective (Maxwell, 2012). The faculties included in the study are Architecture and Design (8.25%), Communications (3.83%), Law (9.32%), Engineering (30.10%), Business (33%) and Health (15.50%).

### 2.2. Measures

*Short Satisfaction Scale towards virtual courses (SVC–S),* which is a unidimensional measure composed of four items. Its psychometric properties are the subject of this study (see Appendix A).

*Academic Satisfaction Scale (ASS)* designed by Lent et al. (2007) in the version of Medrano and Pérez (2010). It is a unidimensional scale composed of eight items with three response alternatives (0 = Never, 1 = Sometimes, 2 = Often, 3 = Always). In relation to validity, it was conducted by exploratory factor analysis with factor loadings above .40 and a percentage of variance explained of 49%. Reliability was obtained by Cronbach's alpha (α = .84). Likewise, for the sample under study, the SSA showed excellent goodness-of-fit indices (CFI = 997; SRMR = .993; RMSEA = .078) and good internal consistency reliability (ω = .90).

### 2.3. SVC-S construction procedure

The construction of the test was carried out in three phases (see Figure 1): Phase 1, called theoretical framework, where the scientific literature on satisfaction was reviewed in specialized texts. This search mechanism made it possible to understand the phenomenon under study.

In phase 2, called test development, the operationalization of the construct was sought through a qualitative approach, which also served as a content test. The proposal made by Ventura-León (2021) was used to systematize the information: (a) Familiarization: all the collected information is exposed in a table to be read and reread (see Table 1); (b) Segmentation, relevant information segments are identified; (c) Categorization, information segments are ordered by simile aspect; (d) Correspondence, it is examined how the items contain the previously generated categories. It is worth noting that this procedure allowed the generation of items related to the aspects of satisfaction. Then, the object of satisfaction, which are the virtual courses, was consigned and generated five items, since the purpose was to build a short measure, i.e., 10 items to less (Ziegler et al., 2014).

Additionally, this process helped to think about the structure and format of the test, which for the purposes of the COVID-19 pandemic was online. Once all this was done, the instrument was applied collectively

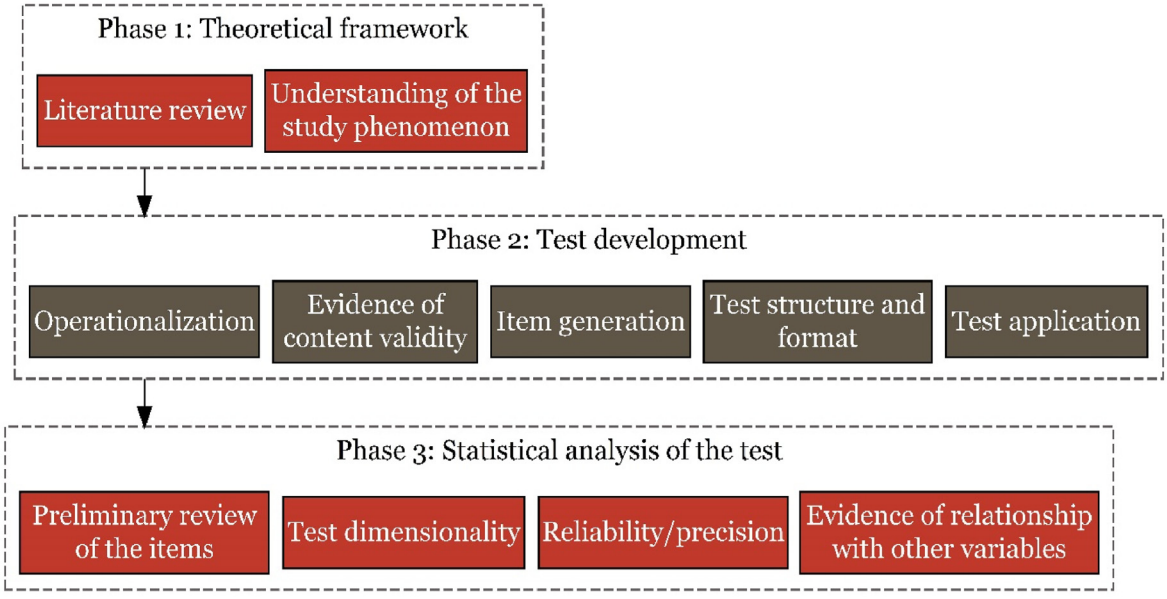


**Figure 1.** SVC-S construction process.

through a virtual form shared through the students' virtual classroom. In this sense, an Internet-based methodology was used (Internet Mediated Research [IMR]); Hoerger and Currell (2011). Previously, the informed consent form was applied, explaining basic aspects such as the objective of the study, anonymity, data processing, etc. In fact, the study was approved by the ethics committee of the university and the guidelines of the Helsinki declaration were considered (World Medical Association, 1964). The virtual form was active from September 27 to October 3, 2021, through the university's official platform, as the study is part of a project of the academic department.

In phase 3, a preliminary review of the items was performed and due to the ordinal nature of the observable variables, a bar graph was preferred. In accordance with international standards, dimensionality (internal structure of the test), reliability/accuracy, and evidence in relation to other variables were examined. This last phase can be seen in more detail in the data analysis section.

All research materials: (a) the database; (b) the R codes; (c) the SVC-S format used can be found in the open access repository OSF: https://osf.io/mwa72/

### 2.4. Data analysis

Statistical analyses were performed with the R programming language in its RStudio space. Specifically, the 'mirt' library was used (Chalmers, 2012), 'ggplot 2' (Wickham et al., 2020), 'tidyverse' (Wickham, 2019), 'semPlot' (Epskamp, 2015) and 'jrt' (Myszkowski, 2021).

Preliminarily, descriptive statistics were calculated through response rates given the ordinal nature of the variables under study. Second, a latent approach such as item response theory (IRT) was used. This approach is complementary to classical test theory (CTT). Specifically, IRT provides parameters for the independent items in the sample under study and provides an information function for the test and the items that allow to observe the accuracy of the test at different levels of the trait measured; instead of a global reliability (Zickar and Broadfoot, 2009). Specifically, a Graded Response Model (GRM) was tested; Samejima (1997) which showed better performance compared to other IRT models (e.g. PCM, GPCM); by presenting lower Bayesian information criterion (BIC; Schwarz, 1978). BIC has proven to be more accurate in polytomous IRT models (Kang et al., 2009). Prior to the use of IRT, its assumptions were reviewed: (a) local independence by the Q3* statistic whose values below 0.20 indicate the presence of this assumption (Christensen et al., 2017); (b) monotonicity; by inspecting the characteristic curve of the categories. Prior to the estimation of the model, we calculated the index *Zh* (Drasgow et al., 1985), a cut-off of ±2.0 was chosen to identify aberrant response patterns (Felt et al., 2017), the revision of outliers is relevant because it can affect the estimation of factorial models (Yuan and Zhong, 2013).

IRT was performed with a two-parameter model (2PL), the discrimination parameter (α) which is the ability of the test to discriminate between people with high and low ability (θ); they are often in the range of -3 to 3; despite this, values greater than 1 can be indicative of high discrimination. The location parameter (β) indicates the value on the θ scale where the person is likely to respond between one response alternative and another. The algorithm to determine the dimensional reduction was MCEM (Monte Carlo EM estimation). The adjustment was performed at two levels: (a) global, through the recommendations of Maydeu-Olivares (2013): Log likelihood, comparative index (CFI ≥.95), Tucker-Lewis index (TLI ≥.95) and root mean square error of approximation (RMSEA ≤.05); (b) local, on items using the RMSEA index which can be considered as a measure of the effect that can range from .05/(k-1) to .089 (Maydeu-Olivares and Joe, 2014) and indicates the magnitude of the difference between the estimated and observed scores in the IRT model. Generalized S-$\chi$2 is not used (Kang and Chen, 2011) as a measure of mismatch because being an inferential process that considers the p-value as a decision criterion requires random sampling (Hirschauer et al., 2020) and is sensitive to sample size (Lin et al., 2013).

Reliability was estimated using the test-item information function, empirical reliability ($r_{xx}$) consisting of *factor score* and model estimates (Du Toit, 2003).

As a final step, convergent and discriminant validity was evaluated using structural equation modeling (SEM). In this regard, the recommendations of Bollen (1989): (a) Specification; (b) Estimation of the parameters, here the WLSMV estimator was considered because it deals with ordinal measures (Brown, 2015); (c) Evaluation of goodness-of-fit considering the cut-offs of Hu and Bentler (1999): SRMR and RMSEA <.08, CFI and TLI >.95. In addition, the recommendations of Fornell and Larcker (1981). In this sense, convergence was evaluated by the average variance extracted (AVE) whose values above .50 are indicators of good convergence. In the case of discrimination, this was done by comparing the square root of the average variance extracted ($\sqrt{\text{AVE}}$) with the correlation between the factors, where $\sqrt{\text{AVE}}$ is expected to be higher than the correlation between the factors. Likewise, a test that measures something similar to the one constructed was selected (Hunsley and Meyer, 2003); in order to find large correlations, but not large enough for conceptual overlap to occur (American Educational Research Association et al., 2019).

## 3. Results

### 3.1. Preliminary analysis

Figure 2 shows the response rates. The highest values are observed to occur in response alternatives 4 (*Neutral*), 5 (*Agree*) and 7 (*Strongly agree*), and the lowest in 1 (*Strongly disagree*) and 2 (*Somewhat disagree*). In addition, there is a tendency to choose high alternatives; although alternative 6 (*Somewhat agree*) shows a decrease in this growth.

### 3.2. Item response theory

Initially, it was examined which of the IRT models best fit the data. This procedure was performed using the BIC which reveals that the GRM (BIC = 34137.68), is the best model when compared to other polytomous models such as PCM (BIC = 35344.86) and GPCM (BIC = 34474.83). The review of the local independence assumptions through *Q3** with the five items revealed a value of 0.30, higher than allowed (*Q3** ≤ 0.20). Then, the residuals matrix was inspected noting that item 3 (*"I like virtual courses"*) had correlations above 0.25 with other items. This implied that item 3 produced noise in the residuals matrix, leading to the decision to remove it. Once with the presence of four items, we proceeded to recalculate the *Q3** index, giving an acceptable value (0.13). Likewise, by inspecting the characteristic curve of the categories, the monotonicity principle is observed (see Figure 3).

Prior to the final analysis, it was decided to check the unusual response patterns through the index *Zh* (Drasgow et al., 1985) taking ±2.0 as the cutoff (Felt et al., 2017). This procedure removed 89 participants whose responses were highly unusual, reducing the data ($n$ = 3080). In this regard, three GRM-2PL models were tested: (a) M1, a model with all five items; (b) M2, a model with four items in the absence of item 3; (c) M3, four items in the absence of response patterns (*Zh* ≤ ± 2.0). The goodness of fit of the models is shown in Table 2.

As M3 demonstrated better performance, we proceeded to examine the model in detail. Table 3 presents the discrimination parameters ($\alpha$) were high for each of the items (i.e., >1.0) and the location values ($\beta$) demonstrate a monotonic increase. Also, item fit measured with the RMSEA index proved to be within acceptable values (RMSEA ≤.089). Thus, the differences between the observed and estimated scores are satisfactory.

### 3.3. Reliability

Reliability was obtained by an empirical coefficient which revealed the presence of good internal consistency at the peak of the trait assessment ($r_{xx}$ = .93). These data are supported by the information

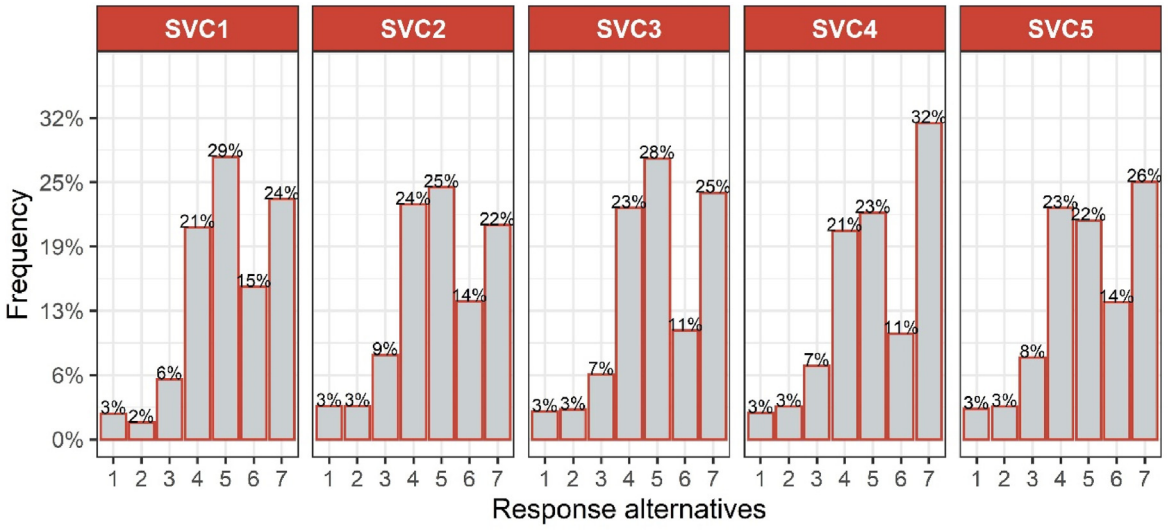


**Figure 2.** SVC-S response rates.

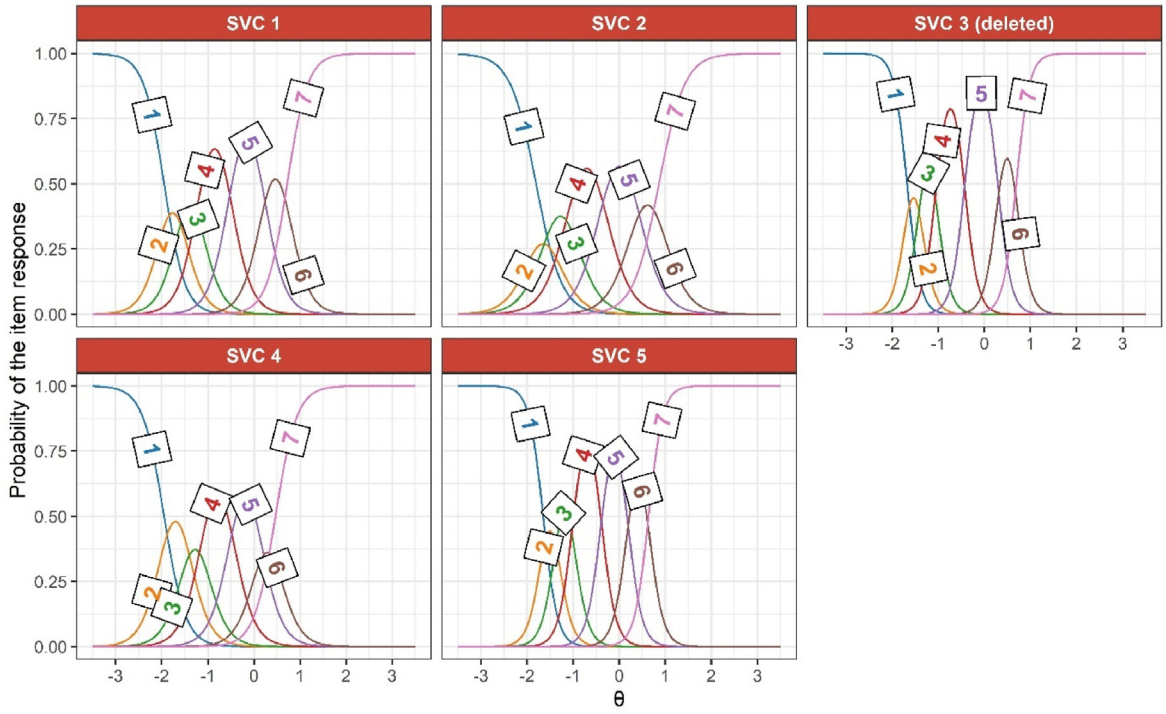


**Figure 3.** Category characteristic curves.

**Table 2.** Fit statistics for the Graded response models 2PL.

| Model | -2LL | AIC | BIC | M2 | df | *p* | RMSEA | CFI |
|---|---|---|---|---|---|---|---|---|
| **M1** | -18567.01 | 37204.01 | 37416.16 | 273.00 | 5 | <.001 | .130 | .987 |
| **M2** | -15975.77 | 32007.55 | 32177.26 | 39.80 | 2 | <.001 | .077 | .997 |
| **M3** | -14587.88 | 29231.76 | 29400.67 | 3.62 | 2 | .164 | .016 | 1.00 |

function and standard error of the test that shows a maximum value of 35.37 ($SE$ = 0.17) when the trait level at θ = -1.40, a situation that suggests that the instrument is more accurate at low levels of the latent trait (Figure 4).

### 3.4. Validity in relation to other variables

Prior to the convergence analysis, the psychometric properties of the SVC-S as a measurement model are examined. Thus, good goodness of fit

**Table 3.** Item statistics for the graduated response model of the SBPS.

| Item | α | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | RMSEA |
|---|---|---|---|---|---|---|---|---|
| **SVC 1** | 4.77 | -1.97 | -1.65 | -1.21 | -0.53 | 0.23 | 0.73 | .085 |
| **SVC2** | 4.37 | -1.87 | -1.51 | -1.07 | -0.36 | 0.36 | 0.85 | .086 |
| **SVC4** | 4.76 | -1.95 | -1.50 | -1.12 | -0.47 | 0.16 | 0.48 | .074 |
| **SVC5** | 7.75 | -1.79 | -1.40 | -1.02 | -0.38 | 0.22 | 0.65 | .082 |

*Note*: α: discrimination parameter; $\beta$: difficulty parameter; RMSEA (measure of effect).

is observed: $\chi 2$ (2) = 4.491; CFI = 1.00; TLI = 1.00; RMSEA = .020; SRMR = .002 and reliability is excellent (ω = .95).

First, the goodness-of-fit measures of the model were reviewed, which to achieve an acceptable value, it was necessary to remove item 7 in the ASS, which had high specific variance, and establish an error correlation between items 5 and 6 of the ASS. This produced excellent goodness-of-fit indices: $\chi 2$ (42) = 872.822; CFI = .993; TLI = .991; RMSEA = .079; SRMR = .030. From this, the relationships in the path diagram were interpreted. In that sense, the AVE revealed a value above .50 for satisfaction with studies (AVE = .83) and academic satisfaction (AVE = .69), demonstrating the convergence of the factors ($r$ = .70). When comparing the $\sqrt{}$AVE with the correlation coefficients they proved to be higher, this is indicative of discriminant validity (Fornell and Larcker, 1981). Details of the correlation in are Figure 5.

## 4. Discussion

Previous health emergencies have already warned of the negative impact of the suspension of educational activities (Cauchemez et al.,

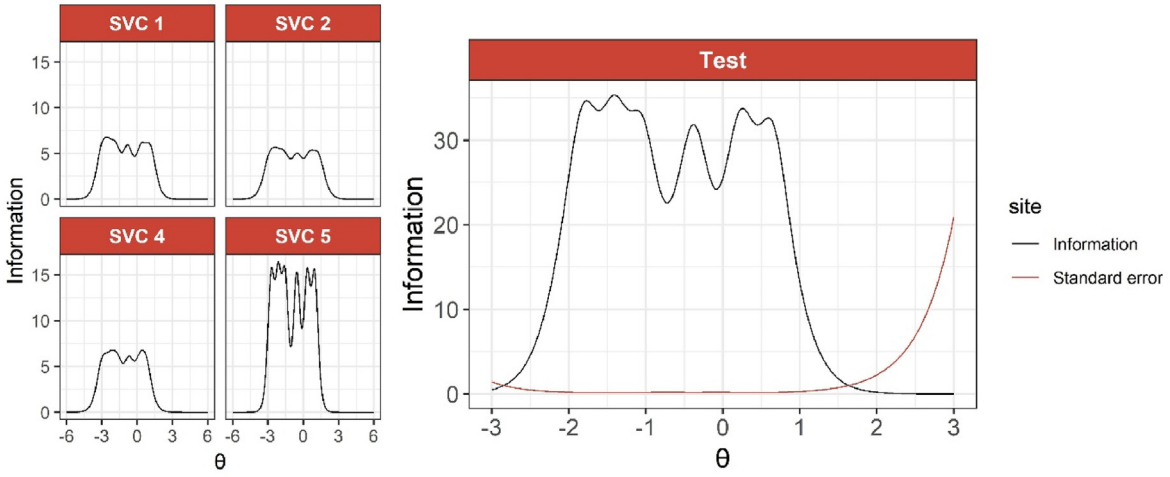


**Figure 4.** Function of the test information and items. Note: In parentheses, the number of the item before the elimination of item 3.

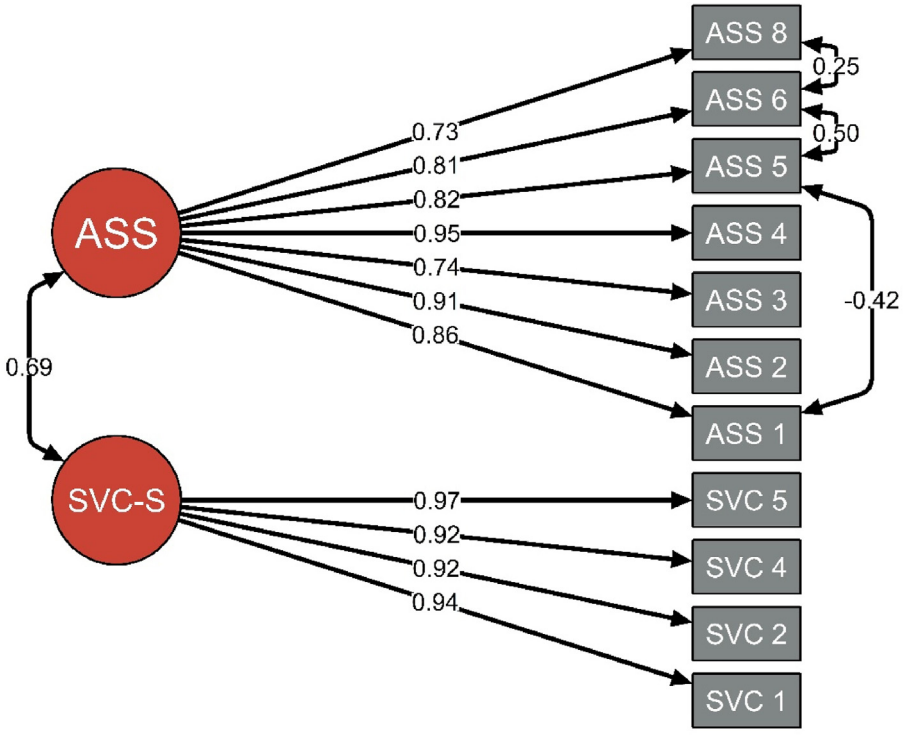


**Figure 5.** Explanatory model between SVC-S and ASS.

2014). In the context of the COVID-19 pandemic, the impact was on millions of students worldwide (Tang et al., 2021). and non-attendance caused abrupt changes in the way higher education was provided, which was the only way to continue education (Hassan et al., 2021; Tejedor et al., 2020) as a COVID-19 containment measure until the development of the vaccine (del Rio and Malani, 2020). In fact, these changes produce a pressure on academic actors such as teachers and students (González-Calvo et al., 2020), possibly because the universities did not have the expertise for virtual courses (Figallo et al., 2020). In this sense, the measurement of satisfaction is timely, even more so when previous studies did not provide psychometric evidence of the instruments used to measure this phenomenon (Jiang et al., 2021), which in some cases were tests *ad hoc* (Chen et al., 2020) or they are too large (Fatani, 2020). In this scenario, the study aimed to develop and validate a short measure of satisfaction towards virtual courses (SVC–S) using modern analysis methods such as the IRT models.

Initially, the instrument is constructed from a qualitative methodology that allows to demonstrating the theoretical approach and correspondence between the items with the theoretical meaning of the test (Ventura-León, 2021). In this sense, satisfaction is a term that comes from an organizational model where the university is understood as a provider of a service and the student as a recipient of that service (Hill et al., 2007; Oliver, 2010; Ragin, 2017; Rai, 2012; Winston and Sommers, 2013). In this sense, satisfaction with virtual courses should be understood as the student's perception of how pleasant the experience of virtual courses is, how they meet their expectations, how interesting they are, and how they become a good alternative for learning.

In relation to content-based evidence, the study did not resort to evaluation by expert judges, but used an alternative procedure based on thematic analysis that stems from the qualitative approach (Ventura-León, 2021). In fact, international standards in psychological and educational testing point out that the content review procedure may include a logical or empirical analysis and may also include expert judgments (American Educational Research Association et al., 2019). Consequently, review by expert judges is not the only way to assess the quality of item content; alternative methods can be considered (Ventura-León, 2021; véase Tabla 1). Furthermore, establishing the suitability of the expert judge is complex because to date there are no clear criteria as to what makes a judge an expert. Therefore, a logical analysis and an inductive procedure were preferred for the review of the content of the items.

On the other hand, the metric was performed with IRT, considering a model that assumes the ordinal nature of the responses (Samejima, 1997). Two parameters such as discrimination and localization were considered. The SVC-S demonstrated good properties in both metrics. Apart from the test and item fit indices, they indicate optimal psychometric properties for the use of the test (Maydeu-Olivares, 2013), within the framework of some explanatory or predictive model (Ato et al., 2013).

An anecdotal issue was the removal of item 3 (*"I like the virtual courses"*) which drastically increased the goodness of fit by providing a high residual relationship with other items; this could reflect that it measures more than just satisfaction; an aspect that breaks the assumption of local independence (Christensen et al., 2017). In fact, this seems to be in accordance with its content, because a taste entails a response based on the sensory and temporal experience of something (Real Academia Española, 2014) which can be difficult to capture. In addition, phrasing has a very broad meaning that excludes relevant aspects of the specific context (Most and Zeidner, 1995). That is, the item does not indicate what exactly they like about the virtual courses. This occurred despite the fact that the item is similar to another satisfaction scale in the academic context (Medrano and Pérez, 2010). This raises further research on the impact of taste questions in the context of satisfaction measurement in university students.

After eliminating item 3, the SVC-S consisted of four items that showed strong measures of fit. Especially, item 4 (*"In general, I feel satisfied with the virtual courses"*); that is, this item would allow distinguishing much better between university students with different levels of satisfaction with virtual courses. In fact, that a single item captures the satisfaction construct sufficiently well is not surprising because the literature provides evidence of something similar in the context of life satisfaction; however, caution is also warranted with these general questions because of their predisposition to measurement error and lower reliability (Jovanović and Lazić, 2020).

Reliability was estimated by the test information function and standard error, which shows that the test as a whole presents a good performance. Performing better for low levels of the latent trait (Chalmers, 2012). However, in local mode, item 4 (*"In general, I feel satisfied with the virtual courses"*) presented a different behavior (spiky distribution) because of high discrimination parameters that could reveal a large sampling variability (Feuerstahler, 2020). However, the evidence of a relationship with another dummy variable as a convergence criterion (Hunsley and Meyer, 2003) demonstrates internal consistency in the answers and prevents conceptual overlap; aspects that have vital importance in the construction of a test (American Educational Research Association et al., 2019).

One of the main strengths of the study was the large sample size. Additionally, another strength was the use of modern psychometric techniques for scale construction, such as the IRT model, and the association of SVC-S scores with ESA scores. Furthermore, the brevity of the SVC-S has advantages for educational and research settings. Thus, the SVC-S can be more easily incorporated into large-scale research, as it would allow for a reduction in the time required to answer questions without a decrease in the information obtained (Ventura-León, 2021). Similarly, it would minimize the burden on the student and allow researchers to develop more parsimonious questionnaires. Also, the SVC-S can be easily integrated into educational practice, where low scores could be a warning sign of the level of individual or group satisfaction with virtual courses. Nevertheless, some limitations of the study should be considered. Despite the large sample size, it was not representative of the Peruvian university population, especially in terms of age and gender, which increases sampling variability (Feuerstahler, 2020). Therefore, the results could not be generalized, making further research necessary. Also, the cross-sectional design is another limitation, which did not allow us to assess test-retest reliability. However, future studies need to use longitudinal designs to examine the stability of scores over time. Finally, the item fit indices were very close to the maximum allowable limit ($\leq$.089); but, estimating accurate scores in the context of such a heterogeneous population (e.g., different schools of study) can be a complex issue; therefore, further study of this aspect is encouraged in future research.

Despite the limitations, the application of the IRT model has provided evidence of the reliability and validity of the SVC-S. Importantly, the analysis showed that the items are unidimensional, locally independent, and reliable with acceptable item fit. Moreover, it showed relationships with another measure of academic satisfaction. These findings suggest future directions for the construction of scales and refinement of items measuring satisfaction in university student population. Thus, the use of measures of satisfaction with virtual courses is suggested to be shorter and more efficient with highly discriminating and internally valid items.

## Declarations

### *Author contribution statement*

José Ventura-León: Conceived and designed the experiments; Wrote the paper.

Tomás Caycho-Rodríguez: Analyzed and interpreted the data; Wrote the paper.

Jency Mamani-Poma, Lucerito Rodriguez-Dominguez, Luciana Cabrera-Toledo: Contributed reagents, materials, analysis tools or data; Wrote the paper.

*Funding statement*

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

*Data availability statement*

Data associated with this study has been deposited at https://osf.io/mwa72/

*Declaration of interest's statement*

The authors declare no conflict of interest.

*Additional information*

No additional information is available for this paper.

## Acknowledgements

Not applicable.

## Appendix A

Short Satisfaction Scale towards virtual courses (SVC–S).
A continuación, encontrarás algunas preguntas acerca de tu **EXPERIENCIA CON LOS CURSOS VIRTUALES**

| Ítems | Totalmente en desacuerdo | Algo en desacuerdo | En desacuerdo | Neutral | De acuerdo | Algo de acuerdo | Totalmente de acuerdo |
|---|---|---|---|---|---|---|---|
| 1. *Los cursos* virtuales son interesantes. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 2. *Los cursos* virtuales cumplen con mis expectativas. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 3. *Los cursos* virtuales me gustan*. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 4. *Los cursos* virtuales son una buena alternativa para el aprendizaje. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 5. En general, me siento satisfecho con los cursos virtuales. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

*Note*. *: Item deleted.

## References

Agarwal, S., Kaushik, J.S., 2020. Student's perception of online learning during COVID pandemic. Indian J. Pediatr. 87 (7), 554.
Aguilera-Hermida, A.P., Quiroga-Garza, A., Gómez-Mendoza, S., Del Río Villanueva, C.A., Avolio Alecchi, B., Avci, D., 2021. Comparison of students' use and acceptance of emergency online learning due to COVID-19 in the USA, Mexico, Peru, and Turkey. Educ. Inf. Technol. 26 (6), 6823–6845.
Almusharraf, N., Khahro, S., 2020. Students satisfaction with online learning experiences during the COVID-19 pandemic. Int. J. Emerg. Technol. Learn. (IJET) 15 (21), 246.
American Educational Research Association, 2019. American Psychological Association, & National Council on Measurement in Education. In: Estándares para Pruebas Educativas y Psicológicas. Estándares para Pruebas Educativas y Psicológicas.
Ato, M., López, J.J., Benavente, A., 2013. Un sistema de clasificación de los diseños de investigación en psicología. An. Psicolog. 29 (3), 1038–1059.
Baber, H., 2020. Determinants of students' perceived learning outcome and satisfaction in online learning during the pandemic of COVID19. J. Educ. E-Learning Res. 7 (3), 285–292.
Basuony, M.A.K., EmadEldeen, R., Farghaly, M., El-Bassiouny, N., Mohamed, E.K.A., 2021. The factors affecting student satisfaction with online education during the COVID-19 pandemic: an empirical study of an emerging Muslim country. J. Islam. Mark. 12 (3), 631–648.
Betts, K., 2009. Lost in translation: importance of effective communication in online education. Online J. Dist. Learn. Adm. 12 (2). https://www.westga.edu/&distance/ojdla/summer122/betts122.html.
Bollen, K.A., 1989. Structural Equations with Latent Variables. John Wiley & Sons.
Brown, T.A., 2015. Confirmatory Factor Analysis for Applied Research. Guilford publications.
Cauchemez, S., Fraser, C., Van Kerkhove, M.D., Donnelly, C.A., Riley, S., Rambaut, A., Enouf, V., van der Werf, S., Ferguson, N.M., 2014. Middle East respiratory syndrome coronavirus: quantification of the extent of the epidemic, surveillance biases, and transmissibility. Lancet Infect. Dis. 14 (1), 50–56.
Chakraborty, P., Mittal, P., Gupta, M.S., Yadav, S., Arora, A., 2021. Opinion of students on online education during the COVID-19 pandemic. Hum. Behav. Emerg. Technol. 3 (3), 357–365.
Chalmers, R.P., 2012. Mirt : a multidimensional item response theory package for the R environment. J. Stat. Software 48 (6), 1–29.
Chen, T., Peng, L., Yin, X., Rong, J., Yang, J., Cong, G., 2020. Analysis of user satisfaction with online education platforms in China during the COVID-19 pandemic. Healthcare 8 (3), 200.
Christensen, K.B., Makransky, G., Horton, M., 2017. Critical values for yen's Q 3 : identification of local dependence in the rasch model using residual correlations. Appl. Psychol. Meas. 41 (3), 178–194.
del Rio, C., Malani, P.N., 2020. COVID-19—new insights on a rapidly changing epidemic. JAMA 323 (14), 1339.
DeMars, C., 2010. Item Response Theory. Oxford University Press.
Drasgow, F., Levine, M.V., Williams, E.A., 1985. Appropriateness measurement with polychotomous item response models and standardized indices. Br. J. Math. Stat. Psychol. 38 (1), 67–86.
Du Toit, M., 2003. IRT from SSI: Bilog-MG, Multilog, Parscale, Testfact. Scientific Software International.
Edelen, M.O., Reeve, B.B., 2007. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual. Life Res. 16 (S1), 5–18.
Embretson, S.E., Reise, S.P., 2000. Item Response Theory for Psychologists. Lawrence Erlbaum Associates, Publishers, Maheah. New Jersey.
Epskamp, S., 2015. semPlot: unified visualizations of structural equation models. Struct. Equ. Model.: A Multidiscip. J. 22 (3), 474–483.
Fatani, T.H., 2020. Student satisfaction with videoconferencing teaching quality during the COVID-19 pandemic. BMC Med. Educ. 20 (1), 396.
Favale, T., Soro, F., Trevisan, M., Drago, I., Mellia, M., 2020. Campus traffic and e-Learning during COVID-19 pandemic. Comput. Network. 176, 107290.
Felt, J.M., Castaneda, R., Tiemensma, J., Depaoli, S., 2017. Using person fit statistics to detect outliers in survey research. Front. Psychol. 8.
Feuerstahler, L.M., 2020. Metric stability in item response models. Multivariate Behav. Res. 1–18.
Figallo, F., González, M.T., Diestra, V., 2020. Perú: Educación superior en el contexto de la pandemia por el COVID-19, 8. Revista de Educación Superior En América Latina. https://bit.ly/3ok72lB.
Fornell, C., Larcker, D.F., 1981. Structural equation models with unobservable variables and measurement error: algebra and statistics. J. Market. Res. 18 (3), 382–388.
George, M.L., 2020. Effective teaching and examination strategies for undergraduate learning during COVID-19 school restrictions. J. Educ. Technol. Syst. 49 (1), 23–48.
Gonçalves, S.P., Sousa, M.J., Pereira, F.S., 2020. Distance learning perceptions from higher education students—the case of Portugal. Educ. Sci. 10 (12), 1–15.

González-Calvo, G., Barba-Martín, R.A., Bores-García, D., Gallego-Lema, V., 2020. Aprender a ser docente sin estar en las aulas: la COVID-19 como amenaza al desarrollo profesional del futuro profesorado. Int. Multidiscipl. J. Soc. Sci.

Hall, T., Connolly, C., Ó Grádaigh, S., Burden, K., Kearney, M., Schuck, S., Bottema, J., Cazemier, G., Hustinx, W., Evens, M., Koenraad, T., Makridou, E., Kosmas, P., 2020. Education in precarious times: a comparative study across six countries to identify design priorities for mobile learning in a pandemic. Inf. Learn. Sci. 121 (5/6), 433–442.

Hambleton, R.K., 1989. Principles and selected applications of item response theory. In: Linn, R.L. (Ed.), Educational Measurement. Macmillan Publishing Co, Inc; American Council on Education, pp. 147–200.

Hambleton, R.K., Swaminathan, H., Rogers, H.J., 1991. Fundamentals of Item Response Theory, 2. Sage.

Hamdan, K.M., Al-Bashaireh, A.M., Zahran, Z., Al-Daghestani, A., AL-Habashneh, S., Shaheen, A.M., 2021. University students' interaction, Internet self-efficacy, self-regulation and satisfaction with online education during pandemic crises of COVID-19 (SARS-CoV-2). Int. J. Educ. Manag. 35 (3), 713–725.

Hassan, S.U.N., Algahtani, F.D., Zrieq, R., Aldhmadi, B.K., Atta, A., Obeidat, R.M., Kadri, A., 2021. Academic self-perception and course satisfaction among university students taking virtual classes during the covid-19 pandemic in the kingdom of saudi-arabia (Ksa). Educ. Sci. 11 (3), 134.

Hill, N., Roche, G., Allen, R., 2007. Customer Satisfaction: the Customer Experience through the Customer's Eyes. The Leadership Factor.

Hirschauer, N., Grüner, S., Mußhoff, O., Becker, C., Jantsch, A., 2020. Can p-values be meaningfully interpreted without random sampling? Stat. Surv. 14, 71–91.

Hoerger, M., Currell, C., 2011. Ethical issues in Internet research. In: APA Handbook of Ethics in Psychology, Vol 2: Practice, Teaching, and Research. American Psychological Association, pp. 385–400.

Hu, L., Bentler, P.M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct. Equ. Model.: A Multidiscip. J. 6 (1), 1–55.

Hunsley, J., Meyer, G.J., 2003. The incremental validity of psychological testing and assessment: conceptual, methodological, and statistical issues. Psychol. Assess. 15 (4), 446–455.

Jiang, H., Islam, A.Y.M.A., Gu, X., Spector, J.M., 2021. Online learning satisfaction in higher education during the COVID-19 pandemic: a regional comparison between Eastern and Western Chinese universities. Educ. Inf. Technol. 26 (6), 6747–6769.

Jovanović, V., Lazić, M., 2020. Is longer always better? A comparison of the validity of single-item versus multiple-item measures of life satisfaction. Appl. Res. Qual. Life 15 (3), 675–692.

Kang, T., Chen, T.T., 2011. Performance of the generalized S-X2 item fit index for the graded response model. Asia Pac. Educ. Rev. 12 (1), 89–96.

Kang, T., Cohen, A.S., Sung, H.-J., 2009. Model selection indices for polytomous items. Appl. Psychol. Meas. 33 (7), 499–518.

Lin, M., Lucas Jr., H.C., Shmueli, G., 2013. Research commentary—too big to fail: large samples and the p-value problem. Inf. Syst. Res. 24 (4), 906–917.

Maxwell, J.A., 2012. Qualitative Research Design: an Interactive Approach. Sage publications.

Maydeu-Olivares, A., 2013. Goodness-of-Fit assessment of item response theory models. Meas.: Interdiscipl. Res. Perspect. 11 (3), 71–101.

Maydeu-Olivares, A., Joe, H., 2014. Assessing approximate fit in categorical data analysis. Multivariate Behav. Res. 49 (4), 305–328.

Medrano, L.A., Pérez, E., 2010. Adaptación de la escala de satisfacción académica a la población universitaria de Córdoba. Summa Psicológica UST 7 (2), 5–14.

Moshagen, M., Erdfelder, E., 2016. A new strategy for testing structural equation models. Struct. Equ. Model.: A Multidiscip. J. 23 (1), 54–60.

Most, R.B., Zeidner, M., 1995. Constructing personality and intelligence instruments. In: International Handbook of Personality and Intelligence. Springer US, pp. 475–503.

Myszkowski, N., 2021. Development of the R library "jrt": automated item response theory procedures for judgment data and their application with the consensual assessment technique. Psychol. Aesthetics, Creativ. Arts 15 (3), 426.

Oliver, R.L., 2010. Satisfaction: A Behavioral Perspective on the Consumer: A Behavioral Perspective on the Consumer. Routledge.

Ragin, D.F., 2017. Health Psychology: an Interdisciplinary Approach. Routledge.

Rai, A.K., 2012. Customer Relationship Management: Concepts and Cases. PHI Learning Pvt. Ltd.

Real Academia Española, 2014. Diccionario de la lengua castellana (23.ªed.). Imprenta Nacional.

Samejima, F., 1997. Graded response model. In: Handbook of Modern Item Response Theory. Springer, pp. 85–100.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2).

Tang, Y.M., Chen, P.C., Law, K.M.Y., Wu, C.H., Lau, Y., Guan, J., He, D., Ho, G.T.S., 2021. Comparative analysis of Student's live online learning readiness during the coronavirus (COVID-19) pandemic in the higher education sector. Comput. Educ. 168, 104211.

Tejedor, S., Cervi, L., Tusa, F., Parola, A., 2020. Educación en tiempos de pandemia: reflexiones de alumnos y profesores sobre la enseñanza virtual universitaria en España, Italia y Ecuador. Revista Latina 78, 1–21.

Ustun, G., 2021. Determining depression and related factors in a society affected by COVID-19 pandemic. Int. J. Soc. Psychiatr. 67 (1), 54–63.

van der Linden, W.J., Hambleton, R.K., 1997. Item response theory: brief history, common models, and extensions. In: Handbook of Modern Item Response Theory. Springer, New York, pp. 1–28.

Ventura-León, J., 2021. Instrumentos breves: un método para validar el contenido de los ítems. Andes Pediatrica 92 (5), 812–813.

Wickham, H., 2019. Tidyverse: Easily Install and Load the 'Tidyverse'. R Package Version 1.3.0. https://cran.r-project.org/web/packages/tidyverse/index.html.

Wickham, H., Chang, W., Henry, L., Pedersen, T.L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., 2020. ggplot 2: create elegant data visualisations using the grammar of graphics (Version 3.3. 0)[Computer software]. Retrievd from Https://CRAN.R-Project. Org/Package= Ggplot2.

Winston, W., Sommers, P.A., 2013. Consumer Satisfaction in Medical Practice. Routledge.

World Medical Association, 1964. Declaración de Helsinki. http://www.conamed.gob.mx/prof_salud/pdf/helsinki.pdf.

Yuan, K.-H., Zhong, X., 2013. Robustness of fit indices to outliers and leverage observations in structural equation modeling. Psychol. Methods 18 (2), 121–136.

Zeng, X., Wang, T., 2021. College student satisfaction with online learning during COVID-19: a review and implications. Int. J. Multidiscip. Perspect. High. Educ. 6 (1), 182–195. https://www.ojed.org/index.php/jimphe/article/view/3502/1483.

Zickar, M.J., Broadfoot, A.A., 2009. The partial revival of a dead horse? Comparing classical test theory and item response theory. Stat. Methodol. Myths and Urban Legends: Doctrine, Verity and Fable in the Organ. Soc. Sci. 37–59.

Ziegler, M., Kemper, C.J., Kruyen, P., 2014. Short scales – five misunderstandings and ways to overcome them. J. Indiv. Differ. 35 (4), 185–189.