(Re-)validation of a standardised test instrument in a different national context

Julie Direnga  
Hamburg University of Technology, Hamburg, Germany  
julie.direnga@tuhh.de

Christian Kautz  
Hamburg University of Technology, Hamburg, Germany  
kautz@tuhh.de

Abstract: The development and application of standardised test instruments such as concept inventories have fostered the discussion on and innovation of teaching and learning. This paper addresses the necessity of validation studies when using a concept inventory in a different national context than the one it was developed in. The revalidation procedure is illustrated using the Concept Assessment Tool for Statics (CATS) in a mixed-methods approach based on (qualitative) student and expert interviews, as well as (quantitative) factor analysis, item response theory, classical test theory, and correlation analysis with exam scores. The results of the quantitative analysis are comparable to prior validation studies performed in the original context. The qualitative analysis raises a few issues but allows for the general conclusion that the CATS may be used and interpreted in the German higher education engineering context as a measure for conceptual understanding of statics.

Introduction

Like any measurement instrument, conceptual tests must measure what they are supposed to measure, and they must do so to a certain degree of accuracy. Whenever a test is applied in a different context than the one for which it was designed, the question should arise whether it still functions as expected. Education researchers are often using tests, for example, to investigate the effectiveness of teaching. Instead of exams, which differ among courses and cohorts, standardised test instruments are often chosen to make use of a common scale. The prefix "standardised" states that the test given to all groups of participants as well as the administration conditions and grading system are the same. Concept inventories (CIs) are standardised tests which focus on understanding of one or several related concepts. In Science, Technology, Engineering and Mathematics (STEM) education research, they are often used to compare different courses or teaching methods with respect to understanding of relevant concepts or as formative assessment to reveal conceptual difficulties on which the instruction should focus. The Force Concept Inventory (FCI), developed in 1992 by Hestenes and Halloun (Hestenes et al., 1992), has probably been the most successful and most widely used CI in physics education research (Engelhardt, 2009). The success of the FCI sparked the development of CIs in other disciplines, predominantly in the US. As of today, more than 150 grants have been awarded by the National Science Foundation (NSF) to projects mentioning the terms "concept inventory" or "concept inventories" in their project titles or abstracts (National Science Foundation, 2018). If possible, existing CIs should thus be used instead of "reinventing the wheel".

The adaptation to a different national context often requires a translation which should be carefully validated. For example, Benegas and Flores (2014) used a Spanish translation of the "Determining and Interpreting Resistive Electric Circuits Concepts Test" (DIRECT) (Engelhardt and Beichner, 2004) that was validated by experts in the field. Mazak et al. (2014) developed a Spanish version of the Concept Assessment Tool for Statics (CATS) (Steif and Dantzler, 2005) and the effect of the translation on performance was investigated in a small-scale control vs. experimental group study. But differences due to national context are not
Conceptual understanding of statics

Drawing forces on separated bodies
Newton’s Third Law
Static equivalence
Roller joint
Pin-in-slot joint
Loads at surfaces with negligible friction
Representing loads at connections
Limits on friction force
Equilibrium

Research Questions

If the German CATS version is administered in the German higher education context, ...

1. ... is the total score a valid interpretation of the level of conceptual understanding of statics?

2. ... do the proposed concept subscales provide valid and reliable information on student understanding of those concepts?

Theoretical Framework

This study is supported by classical and modern test theory which provide a variety of methods (see Methods section). Validity and reliability are central concepts of test theory. Validity addresses the issue of measuring the intended construct, while reliability refers to the degree of accuracy of the measurement. There are several aspects which might pose a threat to validity, for example, if constructs tested by single items are unrelated to the overall construct.
targeted by the test, or if items are systematically misinterpreted. Random misinterpretation of items negatively affects reliability.

The statements "the test is valid" or the term "validity of a test" can be found in many publications. Care must be taken as they falsely suggest that validity is a property of a test. Instead, validity refers to the interpretation of test scores for a predefined purpose. "The test is valid" is a mere shortcut for saying that the interpretation of the test scores for a predefined purpose is valid, if the test was administered in a predefined way. Consequently, the interpretation and the purpose must both be made explicit before any investigation on validity can be commenced (American Educational Research Association et al., 2014). According to Crocker and Algina (1986), validation is "the process by which a test developer or user collects evidence to support the types of inferences that are to be drawn from test scores". Below, we will describe the pieces of evidence that can be collected to support or rebut a validity claim.

Methods

The revalidation is based on the framework proposed for the validation of CI-claims by Jorion et al. (2015) and the procedure for development of a standardised test described by Adams and Wieman (2011). Jorion et al. (2015) suggest to apply classical test theory (CTT) and item response theory (IRT) methods to the entire test to determine problematic items, followed by a structural analysis with problematic items removed. The structure of the instrument shall be first explored with exploratory factor analysis (EFA), followed by an optional confirmatory factor analysis (CFA) or diagnostic classification modelling (DCM). A categorical judgement scheme is provided together with judgement rules to assess the results. Adams and Wieman (2011) additionally suggest correlations with course outcomes such as exams to add evidence that the intended construct is targeted by the test.

These rather quantitative approaches are complemented by qualitative investigations. Adams and Wieman (2011) propose to "[e]stablish topics that are important to teachers" and to "[c]arry out validation interviews with both novices and subject experts on the test questions" (Adams and Wieman, 2011, p. 6). These primarily development-related investigations are also relevant for validation in a different national context. They touch on the aspects of possible differences in terms of course content and learning goals, as well as language, notations, and conventions.

Neither of the mentioned frameworks addresses the special issue of a different national context, but the general goal is the same: making sure that the instrument measures the intended construct. Therefore, most of the suggested methods apply. In addition, a language analysis will be performed, inspecting the differences introduced by the translation and evaluating those in light of the other differences due to national context. An overview over the various analyses is provided in Table 1, stating the purpose and the implementation. A detailed description of the statistical methods cannot be provided within the limited scope of this paper. Please refer to the frameworks of Adams and Wieman (2011) and Jorion et al. (2015) or any standard textbook on test theory (e.g. Crocker and Algina, 1986).

Description of the data

Qualitative data

We interviewed twelve instructors from six different German institutions. As preparation for the interviews, the experts were asked to work through the test. The following are examples of questions from the interview protocol, which served to investigate to what extent experts agree that the test measures the proposed construct:

- Do you expect your students to be able to respond correctly to the items?
- Which concepts do you think does the test address?
- Which central concepts are missing and which ones would you replace or dismiss?
- How do you assess the choice of distractors? Do you recognise the underlying misconcep-
Table 1: Overview of type and purpose of collected evidence for validity

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Purpose</th>
<th>Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Qualitative</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content</td>
<td>Are the concepts tested by the items relevant for local instructors? Do they reflect important learning goals?</td>
<td>Textbook analysis, description of the course, expert interviews</td>
</tr>
<tr>
<td>Language</td>
<td>Did the translation introduce deviations from the original formulations and terms, which changed important aspects of the item?</td>
<td>Item-by-item comparison of translation to original</td>
</tr>
<tr>
<td>Item interpretability</td>
<td>Can local students interpret the items correctly so that the item can measure the intended construct? Do experts evaluate the items as interpretable?</td>
<td>Think-aloud student interviews on selected items, expert interviews</td>
</tr>
<tr>
<td><strong>Quantitative</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTT - item statistics</td>
<td>Determine problematic items in difficulty and discrimination</td>
<td>Apply judgement scheme to difficulty and discrimination indices</td>
</tr>
<tr>
<td>- reliability</td>
<td>Is the scale internally consistent? Do all items add to the reliable measurement of the same overall construct?</td>
<td>Cronbach’s $\alpha$ (of total score and of scores with item $i$ deleted).</td>
</tr>
<tr>
<td>IRT</td>
<td>Do the items fit the assumption that the probability of a correct response to an item is low for students of low ability and high for those of high ability?</td>
<td>Visual inspection of item characteristic curves</td>
</tr>
<tr>
<td>Structural Analysis</td>
<td>How many subscales does the instrument consist of? Which items group together on the same subscale?</td>
<td>Exploratory Factor Analysis</td>
</tr>
<tr>
<td>Exam correlations</td>
<td>Exams measure a related construct. Positive correlation adds evidence to validity.</td>
<td>Pearson’s $r$ between CATS and exams</td>
</tr>
</tbody>
</table>

The interviews were recorded, transcribed, and coded with respect to statements addressing various aspects, such as individual items, the nine CATS concept categories, measurement of expert-like thinking, or expectation of student performance on the CATS. Some of the codes were set a priori, others emerged while reading the transcripts. The range of different responses was very broad, so that not all aspects can be discussed here in detail.

In addition to the experts, 16 students at post-instruction level participated in individual think-aloud interviews. The focus of the interviews was on whether the items are understood as intended. The sample was heterogeneous in terms of the recorded characteristics: gender,
study program, native language, previous education, and their final high school grade. They were all shown the instructions at the beginning of the CATS followed by two to six selected items until the announced time limit of 30 minutes was reached.

Quantitative data
The CATS data was collected in a large introductory mechanics course at Hamburg University of Technology (TUHH) from 2005 to 2016, resulting in a total number of $N_{All} = 4068$ student test data, after removing unserious data with less than nine responses. Students are mostly native German speakers which is also the course language. The test administration was done in class in a paper and pen format. The researchers were not in charge of the course. As course time is sparse, the instructor agreed to provide 45 minutes. The net time on the test was thus 35 minutes, which is substantially less than the proposed 60 minutes. This is expected to have a negative impact on validity. The data indeed consist of many blank responses on the late items, up to 25 % on the last item. They will be graded as incorrect, even though it is unclear if students even got a chance to view the item. We must hence assume that the later items will be overestimated in terms of difficulty. To estimate the extent of this effect, we additionally run the quantitative analysis on a subset of the data, consisting of only those students with no blank responses ($N_{noBlanks} = 2010$). Note that this result will also be biased, though in a different way, because only certain types of students show this kind of test-taking behaviour, i. e. the sample is more homogeneous than the population.

Most of the quantitative results will be compared to the results presented by Jorion et al. (2015). Their sample of $N = 1372$ students consists of aggregated CATS data provided by instructors from about 20 institutions. The aggregation results in high heterogeneity with respect to e. g. pedagogy, courses, or time and method of administration.

Results
In this section, we will present selected pieces of evidence gathered. The qualitative analysis encompasses content, language and item interpretability. The quantitative analysis is mainly based on the framework proposed by Jorion et al. (2015), including CTT, IRT and EFA, and enhanced by a correlation analysis with exams as suggested by Adams and Wieman (2011). The sequence of presentation follows Table 1.

Content
To assess whether the CATS tests concepts which instructors consider relevant for conceptual understanding of statics, a content analysis was conducted based on textbooks, course description and twelve expert interviews with instructors from six different German institutions.

The content analysis showed that the CATS does not address any content which is not introduced by the investigated course, but it does not necessarily address the most important concepts. For example, the correct labelling of forces on a free-body diagram was seen as less important by some experts than the prior step: defining where to separate the bodies. Also, while different types of supports are of course introduced, the focus is different between German and American engineering mechanics instruction in terms of the degree of abstraction in the depicted mechanical systems. US textbooks prefer a more realistic representation, while most German textbooks traditionally use more abstract models, with distinctions made only with respect to the number and type of constraints imposed by the support (as shown in Figure 3 (b)). Consequently, the German students are not familiar with roller and pin-in-slot joints as used in the CATS and thus may have to take an additional mental step to translate the problem into the familiar abstract language.

The expert interviews furthermore revealed that being able to solve any complex statics problem systematically is a central learning goal. In some experts’ opinion, the CATS and the conditions of test administration do not provide the appropriate frame for applying the systematic approach. For instance, one expert (E4) stated that

"The test requires too much intuition for my taste."
When asked what he meant by "intuition", he elaborated:

"Intuition comes when I have solved 20 of such problems or after I have served as a teaching assistant on this topic and have been asked questions which surprised me and I had to think about it myself and question my own... yes, constructed knowledge and connections, then it becomes something like intuition."

If given the appropriate amount of time, however, he would expect his students to solve the questions using the systematic quantitative approach learnt in class. What E4 describes as "intuition", we would call deep or expert-like understanding, which is indeed what the CATS intends to measure. The statement can thus be seen as supporting evidence for validity with the problem being only the time constraint.

The expert interviews were difficult to conduct. As a CI should focus on a limited number of central concepts, the content of an entire course necessarily exceeds the concepts addressed by the CATS. Many of the experts struggled to (1) think in terms of concepts instead of content, (2) see value in a test that does not cover all the course contents, (3) see the different purpose of the CATS compared to an exam, and (4) refrain from criticising the items on a very detailed level. In order to obtain useful and high quality data, it is therefore essential to communicate the purpose and intended scope of the instrument and the focus of the interview. At the same time, the experts' responses must not be biased. Finding the right balance between information and influence is key.

**Language**

The language analysis revealed three main aspects which were affected by the translation: (1) the translated version was making an increased effort to formulate the items precisely and make them "bullet-proof", (2) unfamiliar terms and notations were paraphrased and explained, and (3) conventions were considered.

Nearly all of these adaptations were implemented at the expense of creating more text including more complicated sentences. The construct "reading comprehension ability" becomes increasingly relevant with more complicated phrases, a construct which is not the desired one to measure. In combination with a tighter time limit, this can have a negative impact on the observed performance and thus on validity. As an example for aspect (1), the original question on all 'Pin-in-slot joint'-items ignores that one distractor also includes a couple (see Figure 2). In the translated version, the couple is considered for precision, resulting in a much more

---

2 The correct response is (b).
Figure 3: Different degrees of abstraction. German instruction uses abstract support models, US instruction prefers more realistic representations as starting point to deduct the free-body diagram. ((a) taken from Meriam and Kraige (2008))

complicated sentence\(^3\) (also due to the more complicated German grammar).

The changes were in part influenced by the instructors who had to give permission to administer the test in their courses. Their advice was valuable in terms of learning goals, conventions, and notations, but their tendency to phrase test questions (overly) precisely introduced a higher level of complexity. This tendency is probably culturally influenced, driven by the concern of legal consequences in case of ambiguously phrased exams.

**Item interpretability**

Some of the problems anticipated by the experts, such as interpreting arrows as resultant forces, could be confirmed in the student interviews. Still, most of the selected items were correctly interpreted by the students. Especially the most problematic item in the quantitative analysis, item 20, was most often correctly interpreted. The only misinterpretation that did occur was harmless in that it did not affect the students’ response. The remaining problems to find the correct response were due to lack of conceptual understanding of statics, and therefore a positive indicator for validity.

**Classical Test Theory (CTT)**

According to the scheme proposed by Jorion et al. (2015) (see Table 2), all item difficulties shall lie between 0.2 and 0.8 for an excellent evaluation of item difficulty statistics. For a good evaluation, up to three items may lie outside of this interval. Items 17 and 20 were too difficult in both data sets, “All” and “noBlanks”. As the evaluation of discrimination is determined only by the worst item in the judgement scheme, the result is only average because of the very poorly performing item 20. Removing this item results in good (all item discriminations > 0.1) to excellent (all item discriminations > 0.2) evaluations.

Reliability of a scale is often assessed with Cronbach’s \(\alpha\), which is based on the concept of internal consistency. It assumes that all items measure one single construct. A perfect correlation of 1.0 would indicate that a single item would be sufficient to test the construct. With supposedly nine concepts to test for and a reasonable result for the proposed concept structure (see factor analysis below), a value close to 1.0 is not expected. Furthermore, it should be inspected whether the internal consistency improves or stays the same if an item \(i\) is removed (test if \(\alpha_{-i} \geq \alpha\)). In either case, item \(i\) probably measures a different construct or is badly phrased. We found values slightly smaller than the reference, but still in an acceptable to good range from 0.80 to 0.81. We furthermore found that \(\alpha_{-i} > \alpha\) for item 20.

**Item Response Theory (IRT)**

IRT is a probabilistic test theory. An item’s characteristic curve (see Figure 4 left) shows the estimated probability \(p\) for a correct response depending on the person’s ability (in terms of the construct being measured). Common models are based on the logistic equation which allows to restrict the estimated curves to the interval \([0, 1]\), as required for probabilities. For both our data sets, "All" and "noBlanks", a three-parameter logistic (3PL) model fit the data best.

---

\(^3\)“Welche Richtung hat die Kraft (bzw. welchen Drehsinn hat das Moment), die (das) von der Aussparung auf den Stift im Punkt A ausgeübt wird (werden) [...]”
according to the Akaike information criterion (AIC)\(^4\). The three parameters allow for variation among the items in (1) difficulty, (2) discrimination and (3) guessing probability. Contrary to Jorion et al. (2015), where the two-parameter logistic (2PL) model fit best, we must hence conclude that guessing plays a role in our setting, at least for some items. This may be caused by the tighter time constraint, or it may have cultural reasons.

In the adopted evaluation scheme, the criterion is whether "all items fit the model". Unfortunately, Jorion et al. (2015) leave room for interpretation when this is not the case anymore. We thus define acceptable guessing asymptotes to lie in the interval \([0, 1/3]\), which corresponds to less than two obvious distractors. Furthermore, the probability for high ability persons shall be \(p(4) > 0.8\). Only item 20 does not fit the model in this sense because it is too difficult and must be declared as a problematic item with \(p(4) < 0.8\). This result is in accordance with Jorion et al. (2015) (where item 20 is called Q26). In contrast to the CTT results, item 17 is not problematic when applying IRT. Item 23 has a noticeably high guessing probability, close to but not quite reaching the critical value of one third. Unlike the test information function reported by Jorion et al. (2015), which peaks near Ability = 0, the test information function on the right side of Figure 4 shows that the CATS is most informative for higher abilities in the range between 1 and 2. This indicates that the level of difficulty is slightly too high for the population.

**Factor Analysis**

A factor analysis is applied to investigate the structure of a test scale and check for subscales. The problematic item 20 is omitted. As in Jorion et al. (2015), eight factors were suggested by parallel analysis, regardless of the data set. The exploratory factor analysis was performed with the oblique rotation method 'direct oblimin'. Oblique rotation allows the factors to correlate, which is an appropriate assumption here due to the single scale of conceptual understanding of statics and its high internal consistency. As in Jorion et al. (2015), low factor loadings of less than 0.30 were omitted.

Six factors could be identified to align with the proposed concepts (seven in the noBlanks data set). A concept was determined as identified, if at least two concept-related items and no unrelated items load onto the same factor. The concept 'Loads at surfaces with negligible friction' was only identified in the "noBlanks" data set, the concepts 'Representation of loads

---

\(^4\)The AIC considers not only the goodness of fit, but also punishes high complexity of the model. The introduction of another parameter must hence result in a substantially better fit.
Table 2: Evaluation of psychometric analysis (Jorion et al., 2015, according to scheme suggested in Table 11). E = excellent, G = good, A = average, P = poor, U = unacceptable (did not occur). Numbers in brackets indicate numbers of items which did not comply with the standards. (*corrected error in reference)

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Data set</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>noBlanks</td>
</tr>
<tr>
<td><strong>CTT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Item stats.</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-Difficulty</td>
<td>G .11 to .71</td>
<td>G .11 to .71</td>
</tr>
<tr>
<td>-Discrimination</td>
<td>A .04 to .45</td>
<td>A .08 to .69</td>
</tr>
<tr>
<td>(-Discr. w/o item 20)</td>
<td>G .15 to .45</td>
<td>E .27 to .45</td>
</tr>
<tr>
<td><strong>Total score reliability</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-Cronbach's $\alpha$ total</td>
<td>A .80</td>
<td>G .81</td>
</tr>
<tr>
<td>-Cronbach's $\alpha_{-i}$</td>
<td>G (1)</td>
<td>G (1)</td>
</tr>
<tr>
<td><strong>IRT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Item measures</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-All items fit model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2PL</td>
<td>E (1)</td>
<td>E (1)</td>
</tr>
<tr>
<td>3PL</td>
<td>E (1)</td>
<td>E (1)</td>
</tr>
<tr>
<td><strong>Structural analysis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Exploratory FA</strong></td>
<td>A (7)</td>
<td>G (5)</td>
</tr>
</tbody>
</table>

at connections’ and ‘Equilibrium’ could not be identified by the EFA in either data set. According to the evaluation scheme in Table 2, the EFA results are evaluated as average with seven items that did not load onto the expected factors. If student data with blank responses are removed, the result compares to the reference.

**Exam correlations**

Positive correlations with other instruments measuring a related construct add evidence to the validity of the investigated instrument. As exams are not focused on conceptual understanding to the same extent as the CI, and also cover a broader scope of content, the correlation is not expected to be very high. The correlations (Pearson’s $r$) between the CATS as post-test to statics instruction and the statics exams (midterm and final) were both found to be $r = 0.5 \pm 0.1$, which is comparable to results from Steif and Hansen (2006).

When correlating scores from standardised tests with those from exams, care must be taken to not put too much weight on a single observation, because exams necessarily differ each year. If possible, data from multiple cohorts should be analysed. Depending on local data protection protocols, obtaining and using exam data for research may be difficult. In these cases, gaining validation evidence from correlations with exams may not be possible.

**Conclusions**

The qualitative analysis shows that the content is covered by instruction, and that most (but not all) items are interpretable to students and experts. The selection of concepts was only in part confirmed by the experts to address the most central ones. Applying the quantitative framework by Jorion et al. (2015) reveals that the most problematic item is the same in the German and the US context, but our data show a slightly weaker performance on structure, reliability, and discrimination than the reference. Possible reasons might be that the German
students are less familiar with the representation, that the translated text is more complicated than the original, or that less time was given. Regarding the latter, filtering the data so that no assumptions need to be made on the meaning of a blank response leads to only slightly better psychometric results. Furthermore, Steif and Hansen (2007) found that "[t]here is no noticeable pattern in the variation of mean scores for the remainder of examinees with time above 25 minutes". Even though these results stem from the US context and are not necessarily applicable to the German context, they hint that if there is an effect of giving less time on the test compared to the suggested hour, it may be less serious than expected.

Overall, the collected evidence suggests that the CATS total score can be interpreted as a measure for conceptual understanding of statics for the intended purposes in the German higher education context. For the concept subscales, this interpretability applies only in parts.

Acknowledgements
We thank all experts and students who participated in an interview, and all instructors who allowed us access to their students.

References


