hogrefe

Psychological Test Adaptation and Development (PTAD)

How Papers Are Structured and Why

Hogrefe Publishing GmbH
Merkelstr. 3
37085 Göttingen
Germany

Tel. +49 551 999 50 0
Fax +49 551 999 50 111
publishing@hogrefe.com
www.hogrefe.com

Each submission should adhere as strictly as possible to the following structure. If, for any reason, certain aspects cannot be provided, this should be explained and considered in the limitations and recommendations. **Please see below the table for a detailed explanation.**

Section Content Required	**Specifics for Registered Reports**

Theory / Introduction
 	
A) What is the construct being measured?
 	
• Define each construct measured
• Define possible hierarchy
• Elaborate on score intercorrelations
• Elaborate on nomological net
  – Relations with manifest variables (items)
  – Relations with other constructs
• Derive hypotheses regarding
  – Structural validity
  – Convergent and discriminant validity
• Explain how items were constructed and how they reflect the theory defined above
• Consider consequences of adaptations / translations
 	
B) What is the intended use?
 	
• Define intended use(s) / rule out what it is not intended for
  – Elaborate on necessary data selection contexts
• Derive hypotheses regarding test criterion correlations
  – Pay attention to possible construct overlap, so not only bivariate correlations, use regression, for example, for facet scores
• Delineate requirements for item difficulties
• Delineate requirements for reliability estimates
  – Prognosis vs. status
  – Survey vs. individual assessment
  – Measurement precision	

<table>
<tr>
<td>

C) What is the intended target population?

- Define target population(s)
- Explain how this is adhered to during the studies
- Delineate requirements for item difficulties and content

</td>
<td></td>
</tr>
<tr>
<td>

Methods

- Provide specifics about a possible translation or the adaptation / development undertaken
- Data collection
- Report all measures used in the entire study
- Provide sample information
  - Descriptive statistics
  - Size
  - Composition (e.g., age, gender)
- Justify sample size
- Statistical analyses
  - Define alpha level and if a test is one- or two-tailed
  - Explain which methods match which hypotheses / assumptions and which result is indicative of supporting evidence
  - Structural validity
    - Clearly state whether exploratory (no evidence) or confirmatory
    - EFA vs. CFA
    - Cluster vs. factor mixture modeling
    - Define cutoffs
    - Define what to do in case of misfit
  - Define rules for item selection (e.g., based on loadings, difficulties, variance, content)
  - Report software (version) and packages
  - Provide code that allows a reproduction of analyses

</td>
<td>

- Provide rules for stopping data collection

</td>
</tr>
</table>

Results • Provide all information necessary to evaluate evidence with regard to the assumptions / hypotheses in ABC • **Report all exploratory analyses that were additionally conducted**	
Discussion • Evaluate evidence provided with regard to ABC • Elaborate on limitations • Make clear recommendations how the score(s) can be used	• not required upon initial submission

The structure of a paper in *Psychological Test Adaption and Development* (PTAD) generally follows the typical structure for scientific papers consisting of Introduction / Theory, Methods, Results, and Discussion. We encourage authors to also consider the opportunity to use online supplementary material to support any of these sections.

While the structure in general is pretty much the same as in most other psychological journals, the specific contents of each section are more closely predefined and follow the three lead questions from the "ABC of test construction":

A. What is / are the construct(s) being measured?
B. What are the intended uses?
C. What is the intended target population?

Introduction

The Introduction / Theory section should provide clear-cut answers to all three questions. The logical consequences of these answers will then be the foundation for the ensuing research program.

What Is / Are the Construct(s) Being Measured?

The first part of the theory section should be dedicated to this question. The aim is to clearly define each construct measured in the measurement tool featured. If only one unidimensional construct is assessed, one definition should be provided. However, if there are several test scores or hierarchical structures, this needs to be acknowledged as well. Moreover, in case several constructs are being assessed, the relations assumed between them need to be defined. Finally, a nomological net, meaning

assumed relations with other constructs or observables, needs to be laid out whenever possible. If it is not possible to present a substantive nomological net, then this needs to be acknowledged as well.

The purpose of this exercise is at least twofold. From the perspective of the test users, it will help them interpret the test results with respect to the intended use (see below in **What Are the Intended Uses?**). At the same time, this follows the theoretical guidance by Cronbach and Meehl (1955) regarding construct validity. They stated that each latent variable (i.e., measured construct in most cases) needs to be defined with regard to how it manifests in observable variables and how it relates to other latent variables. In other words, the introduction should allow each reader to understand the nature of the construct(s) measured and the position within a nomological net. Based on this, direct consequences for the kind of validity evidence needed will emerge. The definition provides information about the internal structure of the measure and thus about the assumed structural validity (Loevinger, 1957). Moreover, the nomological net described directly informs possible hypotheses regarding the evidence needed to support convergent and discriminant validity (Campbell & Fiske, 1959; Wehner, Roemer, & Ziegler, 2018). Finally, as this journal specifically calls for submissions featuring test translations or adaptations, it is necessary to elaborate on potential consequences of this procedure and how these are planned to be tested for.

Structural Validity Evidence. Depending on the answer to question A, it is possible that the scores from the featured test reflect a specific theoretical structure. In that case, it will be relevant to test this structure using confirmatory approaches such as structural equation modelling or item response theory. In other cases, a clear structure might not be known, which would justify the use of exploratory approaches. The same might apply when tests are translated or culturally adapted. However, it should be noted that exploratory approaches are only hypotheses generating, not hypotheses testing. Thus, without an additional confirmatory step, using additional data, evidence for structural validity will be limited, which must be acknowledged in the paper.

In cases where only one unidimensional score is expected, necessary steps to provide evidence for this assumption are also necessary. We refer authors to the editorial by Ziegler and Hagemann (2015), who also called for confirmatory approaches. In the case of a misfit, the same paper called for modelling of the misfit in order to gauge its impact on the test score and the relations that are actually interpreted (Heene, Hilbert, Draxler, Ziegler, & Bühner, 2011). It needs to be stressed here that a potential misfit also impacts the choice of reliability estimate as the popular Cronbach's alpha is only suitable for unidimensional items (Cronbach, 1951). Thus, in case of misfit, other estimates such as omega are more suitable (Brunner & Süß, 2005; McNeish, 2018; Revelle & Zinbarg, 2009; Zinbarg, Revelle, Yovel, & Li, 2005).

Convergent Validity Evidence. Evidence for convergent validity, according to Campbell and Fiske (1959), can be obtained by showing that two test scores from different measures which are meant to capture the same construct are correlated more strongly than with scores from tests capturing other constructs. The more modern interpretation often is that the two scores from tests measuring the same construct should be strongly correlated. An even vaguer formulation, which has also found its way into the APA Standards, is to assume a strong correlation between scores from tests which measure the same *or similar* constructs. Schweizer (2012) pointed out that one crucial problem here is the often poor definition of psychological constructs (Michell, 1997, 2001). However, if a clear-cut answer to question A is provided, such a definition should not be found wanting. What remains is the question whether a correlation between scores reflecting similar or identical constructs should both be considered as evidence for convergent validity. Let us consider the example of two intelligence tests. Test A captures verbal reasoning ability and Test B captures figural reasoning ability. Based on existing research it can safely be assumed that the correlation between the verbal and the figural test scores will be around .5 (Schipolowski, Wilhelm, & Schroeders, 2014). Most of us would consider this to be sufficient evidence for convergent validity. However, when gauging the existing evidence for the validity of a test score interpretation, it is also vital to consider the answer process and thus the actual psychological processes that occur when a respondent solves an item (AERA, APA, & NMCE, 2014; Bejar, 1983). Now, whereas the processes (for example inductive and deductive reasoning) might be the same, the content (verbal vs. figural) clearly differs (Schneider & McGrew, 2018). And, importantly, the content is what is driving the difference between different reasoning components (Wilhelm, 2005). Further, when we construct Test A, aiming to measure verbal reasoning, we do not want to measure figural reasoning. Thus, the .5 correlation would not be evidence for convergent validity but should rather be interpreted as discriminant validity evidence (incidentally, a correlation with a test measuring the same construct is now needed and should be higher than .5). Therefore, when choosing the kind of evidence to obtain, the already defined nomological net should serve as a guiding principle. In PTAD we will prefer to see convergent evidence in the stricter sense of Campbell and Fiske. If a more lenient approach, i.e., correlation between scores from tests capturing similar constructs, is taken, a theoretical explanation will be required illustrating why the overlap should be considered due to psychological processes that are in the core of the target construct and not an overlapping construct.

Summing up, based on the answer to question A and the nomological net described, clearly stated hypotheses regarding evidence for convergent validity must be included. Of course, not every paper can include such evidence. However, the lack of it should then be noted as a limitation.

Discriminant Validity Evidence. Whereas convergent validity shows whether a score reflects the intended construct as much as other scores claiming to reflect the same construct, discriminant validity is necessary to show that no other, possibly overlapping, construct is being measured. The example above, using two reasoning tests, shows how strongly the two concepts are intertwined. Excellent examples for the importance of discriminant validity can be found in Mussel (2010) or Credé, Tynan, and Harms (2017) for the constructs of epistemic curiosity and grit, respectively. In both cases, the authors show that once theoretically overlapping constructs or constructs which might be close in the nomological net are considered to be discriminant, the presumed core of the score under scrutiny dissolves into the allegedly discriminant constructs. In other words, when choosing measures to provide discriminant validity evidence, it is important to select measures reflecting close or overlapping constructs. Otherwise, the validity evidence will be very weak (Ziegler, Booth, & Bensch, 2013) and problems of jingle-jangle fallacies (Kelley, 1927) might occur. Again, clear hypotheses need to be formulated.

Item Development. When the construct in question is defined and positioned in a nomological net, item development should be aligned to this information. The paper should report how this was achieved. Ideally, it should be possible to explain how differences in the construct evoke differences during the answering process (Borsboom, Mellenbergh, & Van Heerden, 2003). Moreover, each item should be placeable within the nomological net just defined. In the end, these thoughts ensure content validity.

Consequences of Adaptation. Tests are often translated or adapted to the needs of other populations (e.g., other cultures, age groups). During such processes, several things can happen which might potentially infringe the original purpose of the contained score(s). This is often discussed within the framework of measurement invariance (Borsboom, 2006; Chen, 2008; Fried et al., 2016; Sass, 2011). In fact, without providing evidence regarding measurement invariance with the original version, comparisons of scores are at least highly problematic (Ziegler & Bensch, 2013). Thus, each submission should consider in how far the adaptation or development might have distorted the measure from the original test. Ideally, invariance tests or similar means providing evidence that the adaptation has not resulted in a gross distortion of the relation between definition of the construct and score interpretation should be provided. However, PTAD is not as strict here and does allow submissions without such evidence, for example when the original data are not available. Still, this must be highlighted as a limitation. Basically, in some cases this could mean that the featured measure might only be valuable for research purposes, especially if other validity evidence is also limited.

What Are the Intended Uses?

Each scale development, translation, or adaptation should clearly state the intended uses for the resulting score(s). These can range from research purposes to specific applied uses such as personnel selection or clinical diagnosis. In any case, the intended uses have implications for the setting in which data are collected, criterion-related validity evidence, reliability estimators, and item content. At the same time, it can be advantageous to specifically rule out certain uses for the measure presented. For example, if it is stated that a score should not be used in high-stakes settings, the lack of empirical findings supporting such a use cannot be criticized. In other words, by specifying certain uses and ruling out others, the test constructor(s) set the stage for possible criticism regarding (the lack of) criterion-related validity evidence.

Data Collection Setting. Psychological tests can be used in different settings. On the most general level, we can differentiate between high-stakes settings and low-stakes settings. While test takers may "gain" something for themselves (e.g., a job, a pension, a diagnosis) depending on the test score obtained in a high-stakes setting (Ziegler, Maaß, Griffith, & Gammon, 2015; Ziegler, MacCann, & Roberts, 2011), no such gains are apparent in a low-stakes setting (e.g., research). The consequences of this are mainly visible when it comes to self-reports and the influence of social desirability in low-stakes settings (Bäckström, 2007; Bäckström & Björklund, 2016; Bäckström, Björklund, & Larsson, 2009) or faking / malingering in high-stakes settings (Griffith, Chmielowski, & Yoshita, 2007). To summarize these findings, such response sets or styles often introduce an additional source of systematic variance, thereby potentially inflating internal consistency estimates and intercorrelations of scores obtained with the same method (Ziegler & Bühner, 2009). At the same time, item means and total scores could be distorted. Thus, there can be a strong influence on all kinds of evidence for a score's psychometric quality. Not testing a measure in the intended setting could therefore limit the applicability of the measure. At the same time, it is not always possible to test a measure in the intended setting. This, however, should be clearly stated in the paper and possible limitations, also regarding the uses finally recommended based on the provided evidence, must be discussed.

Criterion-Related Validity Evidence. The implications of the intended uses for criterion-related validity evidence are pretty straightforward. Each submission should include evidence supporting the use of the score(s) and interpretation(s) for the intended applications. Popular examples are correlations between test scores and criteria (e.g., IQ or openness score and school grades). It is also possible to look at mean differences between samples from specific populations. Some measures contain several scores or facets for the construct(s) in question. Here, multivariate analyses are necessary to exemplify the usefulness of each score (Siegling, Petrides, & Martskvishvili,

2015; Ziegler & Bäckström, 2016). In most cases, this will mean providing evidence for incremental validity. For some scores, the intended use is individual assessment (e.g., personnel selection or clinical diagnoses). Here, criterion-related validity evidence should focus on sensitivity and specificity (Kemper, Trapp, Kathmann, Samuel, & Ziegler, 2019). Importantly, clear hypotheses regarding the expected differences or relations must be stated a priori.

Reliability Estimators. Scores from psychological tests are often used for status assessment or prognosis. These different uses have different requirements regarding reliability evidence. For status assessment it is important to showcase the internal consistency of a score. PTAD welcomes estimators that can take structural properties into account (Brunner & Süß, 2005; Revelle & Zinbarg, 2009). If a score is meant for prognoses, reliability evidence should support the stability of a score. Thus, test–retest reliability evidence is vital (Gnambs, 2014, 2015). Again, the use of a score for individual assessment comes along with specific requirements (Sijtsma, 2009; Sijtsma & Emons, 2011). Especially the relation between scale length and reliability must be considered. Here, PTAD strongly encourages the use of the concept of measurement precision (Kemper et al., 2019; Kruyen, Emons, & Sijtsma, 2013a, 2013b, 2014). Thus, the choice of reliability evidence should be justified in accordance with the intended use(s).

Item Content. In some cases, the intended use also has implications for item content. For example, if a score is meant for personnel selection, legislation in some countries does not allow to ask about private matters (e.g., typical vacations). Such matters should be explained. Finally, the intended uses might also have implications for the item difficulties needed.

What Is the Intended Target Population?

The answer to this question has two main implications. First of all, it defines the population from which samples should be drawn for all stages of test construction, adaptation, or development. Second, the answer has implications for the content of the items.

Samples. All samples used in the paper must be drawn from the target population. If, for any reason, this is not the case, explanations regarding possible limitations should be provided. Implications of not using samples from the target population(s) could occur for item means or difficulties and thus item intercorrelations or test score intercorrelations. This should be considered when formulating hypotheses. For example, if a sample used is potentially restricted in variance, correlations between scores could be diminished. This would affect convergent validity evidence negatively. At the same time, discriminant validity evidence could be incorrectly distorted to appear positive.

Item Content. Depending on different characteristics of the sample (e.g., cultural background, age, professional status) the constructs targeted might manifest differently. Consequently, using the same items for different populations can be problematic. Thus, when translating, adapting, or developing a measure, it is vital to demonstrate how the item content relates to characteristics of the target population. For example, pensioners who no longer have a regular job will have problems answering the following conscientiousness item: "I always appear on time at my place of work." Thus, an adaptation of this item to an elderly target population is not straightforward and requires some substantial thinking and possible pre-testing (Ziegler, Kemper, & Lenzner, 2015). The specification of the target population also informs the choice of item difficulties. It is not always wise to simply look for medium difficulties (Ziegler, 2014).

In conclusion, these explanations highlight how answers to the three ABC questions inform and define the plan to evaluate a measure. Each paper should address all of these issues or discuss whether and how the lack of any such evidence decreases generalizability. For registered reports, the same holds true. Moreover, if necessary, papers should explain why certain evidence has not been pursued. At the end of the introduction the reader should have a clear understanding of what is being measured, for what purposes, and in how far the ensuing information supports these claims.

Methods

The typical method section should contain information on how data were collected, **all** measures used, sample demographics and descriptive statistics for the measures used, as well as a section on the statistical analyses (to be) conducted. Whereas most of these parts are pretty straightforward, the sample section and the analyses section come with specific challenges, especially when writing a registered report.

Sample. Whether the paper is a regular submission or a registered report, sample size needs to be justified. In many cases, rules of thumb for manifest (Schönbrodt & Perugini, 2013) or latent correlations (Kretzschmar & Gignac, 2019) will suffice. However, there may be instances where more specific *a priori* power analyses might be required and are really a safeguard against replication failures (Open Science Collaboration, 2015). In such cases, authors often refer to simulations. Here, we just want to point out that, while this is certainly advisable in general, there can be serious problems when the simulations are based on inaccurate assumptions (Albers & Lakens, 2018). Thus, each paper, whether registered report or regular, needs to justify the sample size aimed for or actually acquired.

Statistical Analyses. This section, which is meant to inform the reader about the kind of analyses (to be) conducted and the software used, comes with specific challenges when we adapt or

develop tests. In general, during such processes many decisions need to be made. For example, items might be selected at several stages. Here, the paper needs to accurately inform in detail on decision criteria. In other cases, different reliability estimators could be available (e.g., Cronbach alpha, McDonald omega, or split half). Here, an explanation of why a certain estimator is used should be provided. All of this shows that the choices for the kind of analyses or estimator used must not only be stated but need to be justified! Importantly, this needs to be done in alignment with the answers provided to the three ABC questions. Thus, a kind of protocol matching the pursued evidence (e.g., criterion-related validity evidence) for psychometric quality to analyses (e.g., correlation, regression, or *t*-test) and, importantly, the required result(s) (e.g., expected effect size), needs to be defined. This should be especially fastidious when it comes to structural validity evidence. We will highlight this here using the example of structural equation modeling which is often used to provide such evidence. Here, authors should first clearly define their model of choice and, if possible, alternative models. In a next step, they should state how individual models are evaluated and how different models are being compared. Here, I explicitly refer all authors to literature which suggests a more differentiated approach to standard cutoffs for indices such as RMSEA or CFI (e.g., Greiff & Heene, 2017; Heene et al., 2011). Importantly, authors also need to explain what they plan to do (for a registered report) or justify what they did (for a regular submission) when the preferred model did not fit the data. Options might include to delete items, to add correlated residuals, to add crossloadings, or additional latent variables to name just a few. In a registered report, these choices must be made clear. In all submissions, the consequence of such choices, for example regarding content validity or unidimensionality (Ziegler & Hagemann, 2015), need to be considered and stated.

Authors should not forget that this section is vital when it comes to showcasing how reliability and validity evidence was obtained and how trustworthy it is. Any deviations from the planned procedure must be explained in a registered report and possible limitations added. In regular submissions, the same holds true when there are deviations from the assumptions laid out in the introduction.

Results

This section should contain all information necessary to evaluate whether the score(s) from the measure presented in the paper actually measure the intended construct and are useful in the intended way. In registered reports, the kind of information that is planned to be reported can be portrayed. As before, the option of using online supplementary material should be evaluated.

In general, we require that authors share code, data, and material when submitting. If code, data, or materials cannot be shared for any reason, this must be clearly explained in the manuscript. The general premise here is that any researcher should be able to reproduce the central results of the study without contacting the original authors. This requires open code and open data. Furthermore, it should in principle be possible to replicate the study in an independent sample. This requires open material (i.e., items, instructions, study-set up) or a reference to the source of the materials.

Discussion

Within this section, the evidence obtained should be evaluated with regard to the requirements formulated in the introduction. As a result, clear recommendations should be listed. This refers to whether the measured score(s) can be interpreted as intended.

These, in places detailed, elaborations should be understood as a kind of template. Submissions will be expected to follow the structure laid out here and to provide the information outlined (or explain the lack of it). There will be a formal check of each submission and divergences from this template may result in the paper being send back with a request for closer alignment with the template. Please keep in mind that in doing this we aim to help both readers and reviewers – and also our authors because the chances of a fast peer review process and of the paper being cited later on should improve substantially!

At this point, I would once again like to highlight the option of submitting a registered report. Planning a test translation or adaptation will, we hope, be facilitated by following these guidelines – the template can be used in the sense of a checklist. Moreover, the opportunity to obtain timely feedback, before data are being collected, should allow us to weed out problems or even critical flaws that otherwise could not be undone later. This in turn should bolster the paper's final quality.

Now all that remains for me to say is that I look forward to the start of the new journal – and to reading YOUR paper!

Matthias Ziegler

Humboldt-Universität zu Berlin, Germany

Editor-in-Chief of *Psychological Test Adaptation and Development*

November 2019