Förster, Manuel; Zlatkin-Troitschanskaia, Olga; Brückner, Sebastian; Hiber, Jochen

Johannes Gutenberg-University of Mainz
Department 03 Law, Management and Economics
Chair of Business and Economics Education
55099 Mainz
Germany

Corresponding E-Mail: manuel.foerster@uni-mainz.de

Adapting and Validating the Test of Understanding in College Economics to Assess the Economic Knowledge and Understanding of Students in Germany

1. Research Aims and Questions

In Europe and particularly in Germany, cross-institutional assessment of students' learning in higher education is becoming an increasingly important research area. This is especially true for the field of business and economics, which is the most popular field of study among beginning and advanced students (Federal Statistical Office, 2012). Nevertheless, the field of business and economics still lacks a German-language test instrument that meets academic requirements to assess economic knowledge (Kuhn & Zlatkin-Troitschanskaia, 2011). Previous research approaches have focused on target groups outside of higher education (Nickolaus, 2011). To close the research gap, the WiwiKom project[1] is developing a German language test instrument to measure the economic knowledge of business and economics students in Germany. This will be done by adapting and validating international assessment instruments and merging them into one German test instrument. The adapted test is also meant to enable international comparative studies focusing on, for instance, higher education students in Germany and the U.S. and comparing the economic knowledge they acquire during their studies.

So far in the WiwiKom project, we have adapted and validated the Test of Understanding in College Economics (TUCE) created by the U.S. Council for Economic Education (CEE; Walstad, Watts, & Rebeck, 2007). The TUCE is an internationally approved testing instrument to assess economic knowledge in higher education. In this paper, we present preliminary findings from the adaptation and validation process. Based on this evidence, we explore the extent to which the TUCE can be used

[1] WiwiKom is the abbreviation of a project conducted in Germany called 'Modeling and measuring competencies in business and economics among students and graduates by adapting and further developing existing American and Latin-American measuring instruments'. For more information, see e.g. Förster, Zlatkin-Troitschanskaia, Brückner et al. 2013; http://www.wiwi-kompetenz.de/eng/index.php.

to assess the economic knowledge of undergraduates in the field of business and economics in Germany. First, we explain the criteria we had to meet to adapt and validate the TUCE for the German context. Then, we give an overview of the analyses conducted and preliminary results, from which we draw conclusions about the validity of the adapted test.

2. Theoretical Criteria for Validating Adapted Tests

Adapting a psychological test is a complex and multifaceted task (e.g., Beck & Krumm, 1991). Although the International Test Commission has issued Test Adaptation Guidelines (TAG) to ensure a high quality of test adaptations (Coyne, 2000; Hambleton, 2001), they provide only a rough orientation and need to be specified substantially with regard to the content of the respective project. In WiwiKom, the TAG were used as a methodological basis for the translation and adaptation process. The TAG are briefly described in the following.

The TAG already have been broadly used in test adaptation practice. They have been the basis for the adaptations, for example, of the third Trends in International Mathematics and Science Study (TIMSS) and the Program for International Student Assessment (PISA) (Grisay 2003, Hambleton 2001);however, the guidelines were specified in quite different ways in these studies. The TAG are divided into four sections: (1) The guidelines concerning "context" are meant to ensure that assessment of knowledge of different populations target the same aspects of theoretical constructs as well as to minimize cultural influences irrelevant to the assessment. (2) The guidelines on "test development and adaptation" focus on issues related to translation, data collection, and statistical analyses. (3) The guidelines on "administration" deal with testing procedures and issues arising when subjects have different language and cultural backgrounds. (4) The guidelines on "documentation/score interpretations" emphasize the importance of documenting the test and changes made during the adaptation process to ensure validity and avoid diagnostic misinterpretation.

Test validity, in particular, can be evaluated using the TAG as well as the American Educational Research Association (AERA) Standards. In WiwiKom, the TUCE was validated based on the AERA Standards, which require evidence from five categories, including test content, response processes, internal structure, relations to other variables, and consequences of testing (AERA, 2004, pp. 11-17).

1) Evidence based on test content

Test content is analyzed to determine how accurately it represents theoretical constructs (AERA, 2004, p. 11). The accuracy of representation can be determined through logical or empirical analysis (AERA, 2004, p.11). Another possibility is to ask experts to evaluate how the test content relates to content of the respective field of study. Thus, the content of the TUCE can be evaluated by experts from the respective subject areas of economics. This analysis is particularly important when a test is administered in a new educational context where the educational system and the curriculum might differ from the original target context. Existing curricular differences must be taken into account when adapting and validating a test to enable valid assessment and comparison among countries. In WiwiKom, we had to ensure that the construct of economic knowledge would be conceptualized and understood in a similar way in the U.S. and in Germany (AERA, 2004, p.12). This criterion of the AERA Standards corresponds with the criterion of "context" in the TAG.

2) Evidence based on response processes

Subjects' response processes are analyzed to determine whether there is a good fit between theoretical constructs and the observed item response processes. Subjects' individual response strategies are examined, and items can be revised if they are repeatedly misunderstood. When validating the TUCE, we had to ensure that item responses were indeed based on the intended cognitive processes and test solving strategies and that, in contrast, undesired test-taking and guessing strategies did not result in correct responses (AERA, 2004, pp.12-13).

3) Evidence based on internal structure

The internal structure of a test is analyzed to determine the extent to which a construct is coherently represented by the relations between single items or various parts of a test. How the internal structure is analyzed and how the results are interpreted depends on the aim of the test, that is, on the initially assumed structure. For instance, a one-dimensional construct or test is expected to have rather homogenous items (AERA, 2004, p.13). For the TUCE, we assumed that microeconomics and macroeconomics were clearly correlated, but still separate, dimensions.

4) Evidence based on relations to other variables

Another validity criterion lies in the relationship of the test results to other external variables analyzed according to the relations in a nomological network (AERA, 2004, pp. 13-16; Cronbach & Meehl, 1955). External variables may be personal or group-related variables, and the evidence may reveal a convergent or discriminant relationship between the construct and the respective variable. For instance, according to the expert-novice paradigm (Ericsson, 2008), we assumed that students who majored in business and economics would score higher on the economic knowledge test than students who minored in business and economics or studied subjects unrelated to business and economics.

5) Evidence based on the consequences of testing

The consequences of testing refer to the conclusions drawn from the test scores, which allow evaluation of individual students or groups of students. For the TUCE, this meant primarily that we had to document sufficiently the findings from the test development process. We had to provide a manual that would explain the possible uses of the test so that those applying the test would not draw inadequate conclusions or apply the test in an inappropriate way (AERA, 2004, pp.16-17).

3. Adaptation and Validation Process of the TUCE and Preliminary Results

In the following, we present the adaptation and validation process and highlight the results indicating whether the TUCE is a valid tool for assessing students' economic knowledge in higher education in Germany. We have sorted the analyses according to the above validity criteria although some of the analyses can be attributed to more than one validity criterion.

Translation of the TUCE and linguistic and cultural adaptation

First, the TUCE was translated into German by professional translators specialized in economics to ensure that the translation was of very high quality. Another economic knowledge test, the Test of Economic Literacy (Soper & Walstad, 1987), had already been translated into German by Beck and Krumm (1991). Their approach was to translate word by word and sentence by sentence, but they reported unavoidable grammatical and lexicographical problems. Certain English sentence structures were almost impossible to translate on a word-by-word basis. For certain English terms, there was no exact German translation; hence, the translation had to convey the meaning as closely as possible. The problem of equivalence, that is, how closely a translation must adhere to the original, has been discussed extensively in translation studies, for instance, by Koller (1979) and Albrecht (1990). In more recent years, translation scholars have turned away from equivalence-based translation models and towards target-oriented approaches, such as the functionalist and action-oriented translation models by Reiß and Vermeer (2013/1984). These approaches take into account the linguistic and cultural characteristics of the source and target texts and today are considered more promising for producing high-quality translations. Hence, they also were used in the WiwiKom project. First, the source texts were translated with annotations by a certified translation service provider specialized in the field. Linguistic supervision and quality assurance was provided by the English department of the Faculty of Translation Studies, Linguistics, and Cultural Studies of the Johannes Gutenberg-University Mainz. The entire translation process met the highest academic quality standards. For instance, terminology management was used to optimize consistency and proofreading was done by two professionals to assure quality. Furthermore, a translation workshop was conducted during which the developers of the TUCE were interviewed on specific translation problems. Subsequently, the translation was revised after several feedback sessions during which experts from business and economic studies reviewed the test and reported on any field-related shortcomings of the content. In further workshops, the experts from economic and translations studies came together to revise the translation and validate a final version for use in the field (see D1-D3 of the TAG; Hambleton, 2001). Thanks to this comprehensive process, only few items required further cultural adaptation. For example, one item description included a crop of oranges that had been destroyed by a hurricane, which was adapted to the German context as a crop of apples that had been destroyed by a severe winter frost.

Validation of the TUCE content

To ensure content validity, or curricular validity, the test items were submitted to curricular analyses as well as to an online ratingconducted by experts in cooperation with the German Higher Education Information System (HIS). We analyzed the curricula of 96 degree courses at 40 universities and 24 universities of applied sciences in Germany, including the curricula of the largest business and economics faculties at higher education institutions[2] in Germany and of those faculties that participated in the WiwiKom survey. We examined study program descriptions and module manuals to determine which content constituted a core curriculum, that is, which content was taught at most or all higher education institutions in Germany. Then, we verified whether the content of the test items were part of the curricula of universities and universities of applied sciences in Germany. Overall, results from the curricular analyses showed that the economic content areas of microeconomics and macroeconomics surveyed in the WiwiKom project were taught at most universities and universities of applied sciences in Germany, and the content of the TUCE covered a large part of the core curriculum in economic studies. During the online rating, the TUCE items were further evaluated by 16 professors and lecturers in economics from higher education institutions in Germany. This online rating by experts not only provided a cross-validation of the curricular analyses, but also served to assess additional aspects that were highly relevant for content validation, such as item difficulty. The experts rated the curricular relevance and the item difficulty and gave a general judgment on each item. The online questionnaire consisted of closed-ended rating items on a seven-point Likert scale and provided an area for additional feedback on the items. Overall, content validity was confirmed for all 60 items of the TUCE. The core curriculum of study programs in economics seems to be similar in Germany and the U.S., suggesting that the construct of economic knowledge is understood in a very similar way in both countries. These analyses were further complemented by recurrent expert interviews with lecturers in economics. We discussed problematic items with the lecturers at different points in the process, such as after the translation, after the curricular analyses,

[2] In Germany, the two common types of higher education institutions are universities and universities of applied sciences. Universities mainly aim to provide academic education, while universities of applied sciences are more practically oriented.

and after the online expert rating. The discussions focused on whether the item content was correct and whether the distractors were adequate. If necessary, we looked for alternative phrasings or field-specific terms that would ensure accuracy and increase comprehensibility of the items while altering as little as possible the meaning and focus of the items. All 60 TUCE items were successfully adapted for the German test version, which we attributed to a high degree of cultural comparability between Germany and the U.S.

Validation of the item response processes

The item response processes were validated through standardized cognitive interviews with students. On two measuring dates, students from business and economic degree courses were interviewed during the item response process and immediately after it. While responding to items, the concurrent think-aloud method was applied, and subsequently, it was complemented by a retrospective interview with targeted, standardized questions (Ericsson & Simon, 1993). Due to the strenuous nature of this method, we had to limit interview time to 1.5 hours, and each subject was given no more than 20 items to respond to. Nevertheless, each item was eventually answered by at least eight different students. The interviews on the first measuring date were used to improve "item clarity" (Leighton, Heffernan, Cor, Gokiert & Cui, 2011); they helped us identify general misunderstandings and revise unsuitable phrasing. This analysis focused on problems of graphic representation or typographical and grammatical errors, which made the items more difficult to understand. In some sentences, translation and adaptation led to syntactical changes and created ambiguity. The cognitive interviews also provided an indication of how long the test would take to complete. In this regard, we found significant differences depending on the students' previous knowledge. Thanks to the cognitive interviews, we were able to take this time component into account when designing the test booklets, which was an important point, since the WiwiKom survey was conducted during regular university classes and testing time was limited to one hour. The interviews on the second measuring date were used to analyze the item response processes (Leighton et al., 2011). They served to assess the thought processes that took place while responding to items, which could be used later to interpret test scores. These cognitive interviews can provide further indications of the students' field-specific knowledge structures (Brückner 2013). Currently, we are analyzing the extent to which primarily construct-relevant thought processes resulted in correct responses and whether there was interference from construct-irrelevant test-taking and guessing strategies.

Validation of the internal structure and relations to other variables

The German version of the TUCE was tested empirically for the first time in a major field survey with 878 students. The survey was conducted at 23 higher education institutions in Germany, with 84% of the subjects coming from universities and 16% from universities of applied sciences. The share of female students was 46%. With regard to the study progress, students were on average in their 3rd semester when they participated in the survey. Due to the limited testing time and in order to control for serial position effects, we used a complex multiple matrix design (Frey, Hartig & Rupp 2009). As a consequence, each student was presented with 20 to 30 TUCE items.[3]

Based on this sample, we examined the internal structure of the test to see whether knowledge of microeconomics and macroeconomics represented two distinct dimensions or only one dimension of economic knowledge. Furthermore, we analyzed whether economic knowledge increased according to expectations over the course of bachelor studies and whether there were differences in the knowledge scores of different groups, which had already been confirmed for other testing methods in economics in the German language. We examined whether such differences in test scores were due to influences of other variables, for instance, gender, an economics class attended by the students, or a commercial vocational training completed prior to studies. Such analyses of the relations to other variables provide indications for the convergent or discriminant validity of the test. Our hypotheses were tested using analyses from classical test theory and item response theory. Preliminary results showed that the two dimensions of microeconomics and macroeconomics are clearly correlated but definitely separate, as the factorial analysis showed. Furthermore, regressions on the latent knowledge scores showed that attending a class on microeconomics or macroeconomics had a positive effect on the total score. As expected, the test has convergent validity, as it provided a good measure of the positive effect of previous knowledge and economics classes on economic knowledge. Unfortunately, the students' gender had a strong impact on the test score as well. On both the original American test and the adapted German test, male students performed better than female students when other relevant influence factors were controlled. This might compromise the fairness and validity of the test.

4. Conclusion

Overall, our results show that the TUCE allows valid conclusions to be drawn about how students in higher education in Germany respond to items on economic knowledge. Thus, we can use the TUCE results to draw general preliminary conclusions about the economic knowledge of undergraduate students. In the presentation, we will discuss in greater detail the statistical analyses for the German survey.

[3] The complex multiple matrix design, including 41 questionnaires, was indeed necessary, since the survey consisted not only of the TUCE, but also of other tests assessing business knowledge. Altogether, 220 Items were tested in this survey.

The next step will be to determine the requirements for international comparisons among countries, for instance, between Germany and the U.S. A feasibility study will need to compare in particular the empirical measurement models used in both countries. From the models, it will be possible to draw conclusions about the degree of similarity of the underlying latent constructs in both countries. The TUCE data from Germany and the U.S. currently are being prepared for comparison. Subsequently, various confirmatory factor models and Rasch models can be applied to analyze the data from both countries with regard to different kinds of measurement invariance. The aim will be to conceptualize a joint measurement model which enables estimations, if possible, of unbiased knowledge scores in both countries. Based on the knowledge scores, it will be possible to determine differences and similarities between Germany and the U.S. Then, further questions can be answered, such as to what extent students' economic knowledge differs between the two countries, and whether there are structural differences between the countries with regard to the dimensions of microeconomics and macroeconomics or to certain groups, such as male and female students.

References

Albrecht, J. (1990). Invarianz, Äquivalenz, Adäquatheit. [Invariance, Equivalence, Adequacy.] In: Arntz, Reiner/Thome,Gisela (Eds.): Übersetzungswissenschaft: Ergebnisse und Perspektiven. Tübingen: Narr. S. 71- 81.

American Education Research Association (AERA), American Psychological Association & National Council on Measurement in Education (2004). Standards for educational and psychological testing (revised ed.). Washington, DC: American Educational Research Association.

Beck, K. & Krumm, V. (1991). Economic Literacy in German Speaking Countries and the United States. First Steps to a Comparative Study. Economia, 1(1), 17–23.

Brückner, S. (2013). Construct-irrelevant mental processes in university students' responding to economic test items: Using symmetry based on verbal reports to establish the validity of test score interpretations. Brunswik Society Newsletter, 28.

Coyne, I. (2000). ITC Test Adaptation Guidelines. April 21, Version. International Test Commission. Verfügbar unter: http://www.intestcom.org/test_adaptation.htm [23.08.2013].

Cronbach, L. J. & Meehl, P.E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Ericsson, K. A. (2008). Deliberate practice and acquisition of expert performance: A general overview. Academic Emergency Medicine, 15, 988–994.

Ericsson, K. A. & Simon, H. A. (1993). Protocol analysis: Verbal reports as data. (2nd revised ed.). Cambridge: The MIT Press.

Federal Statistical Office (2012). Studierende an Hochschulen – Wintersemester 2011/ 2012 [Students at Universities – Winter Term 2011/2012]. Wiesbaden, Germany: Statistisches Bundesamt.

Förster, M., Zlatkin-Troitschanskaia, O., Brückner, S. & Hansen, M. (2013). WiwiKom – Modeling and Measuring Competencies in Business and Economics among Students and Graduates by Adapting and Further Developing Existing American and Mexican Measuring Instruments. In Blömeke, S. & Zlatkin-Troitschanskaia, O. (Eds.) The German funding initiative "Modeling and Measuring Competencies in Higher Education" (KoKoHs Working Papers, 3). Berlin & Mainz: Humboldt University & Johannes Gutenberg University.

Frey, A., Hartig, J. & Rupp, A. (2009). Booklet Designs in Large-Scale Assessments of Student Achievement: Theory and Practice. Educational Measurement: Issues and Practice, 28, 39-53.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. Language Testing, 20, 225–240.

Hambleton, R. K. (2001). The Next Generation of the ITC Test Translation and Adaption Guidelines. European Journal of Psychological Assessment, 17, 164–172.

Koller, W. (1979): Einführung in die Übersetzungswissenschaft. [Introduction to Translation Studies.] Heidelberg: Quelle und Mayer.

Kuhn, C. & Zlatkin-Troitschanskaia, O. (2011). Assessment of Competencies among University Students and Graduates – Analyzing the State of Research and Perspectives. Johannes Gutenberg University Mainz: Arbeitspapiere Wirtschaftspädagogik [working paper: business education], 59.

Leighton, J. P., Heffernan, C., Cor, M. K., Gokiert, R. J. & Cui, Y. (2011). An Experimental Test of Student Verbal Reports and Teacher Evaluations as a Source of Validity Evidence for Test Development, Applied Measurement in Education, 24(4), 324-348.

Nickolaus, R. (2011). Kompetenzmessung und Prüfungen in der beruflichen Bildung (Competence Assessment and Exams in Vocational Education). Zeitschrift für Berufs- und Wirtschaftspädagogik, 107(2), 161-173.

Reiß, K. & Vermeer H. J. (2013/1984). Towards a General Theory of Translational Action: Skopos Theory Explained. Manchester: St. Jerome.

Soper, J. C. & Walstad, W. B. (1987). Test of Economic Literacy: Second Edition Examiner's Manual. New York: Joint Council on Economic Education.

Walstad, W. B., Watts, M. & Rebeck, K. (2007). Test of understanding in college economics: Examiner's manual (4th ed.). New York NY: National Council on Economic Education.