Pretest-posttest measurement of the economic knowledge of undergraduates – Estimating guessing effects

Discussion Paper for the AEA Annual Meeting 2018 – Philadelphia

William Walstad[1]; Schmidt, Susanne[2]; Zlatkin-Troitschanskaia[2], Olga; Happ, Roland[2]

Institutions:

[1] University of Nebraska–Lincoln
[2] Johannes Gutenberg University Mainz

Abstract:

Learning economics is often discussed in terms of students' end-of-course achievements as indicated by a test score or course grade. This stock of knowledge provides information on how much students know at a given time. However, not only the learning outcome is important but also the change in patterns over time. In this study, we measured students' economics knowledge and understanding at the beginning (pretest) and at the end (posttest) of their first year at university. In this value-added approach, learning is measured as the change of knowledge and understanding over a period of time. To measure learning, we administered items from the TEL (Walstad, Rebeck, & Butters, 2013) and from the TUCE (Walstad, Watts, & Rebeck, 2007). To ensure that all participants have the same entry conditions to the test, we only sampled students attending an economics introductory course covering economic principles as well as basics of microeconomics and macroeconomics. Thus, undergraduate students of business and economics were assessed over the course of their first year. To investigate the change of economic knowledge and understanding, we followed Walstad and Wagner's (2016) approach to analyze four types of learning (positive, negative, retained, and zero learning). As pointed out by Smith and Wagner (2017), measuring these four types in the context of multiple-choice items also involves taking a closer look at guessing. Students could give a right answer because they know the correct answer or they just guess the right solution. Considering this, the estimates for the four types of economic learning were adjusted to the expected number of correct answers. The presentation shows that the approaches by Walstad and Wagner (2016) and Smith and Wagner (2017) have given new insights into learning patterns and analyzing change of knowledge and understanding, which is important for researchers and lecturers not only in economics education.

Key Words:

Pretest-posttest-design, economic knowledge and understanding, Test of Economic Literacy, Test of Understanding College Economics, Guessing, Multiple-Choice Items

JEL Codes: A20, A22, A23

Relevance

In order to evaluate learning progress in university courses, pretest-posttest designs are often used (see Walstad & Wagner, 2016; Happ, Zlatkin-Troitschanskaia, & Schmidt, 2016). Before course instruction begins, students are given a multiple choice test as a pretest to measure initial knowledge and understanding of a subject. For each multiple choice item, students select the one correct answer from the stated alternatives (usually four or five). The correct responses to items are totaled to produce a pretest score for students. Then, after a time period (e.g., end of a 15-week semester), the same multiple choice test is administered to the same students to assess their final knowledge and understanding. The correct responses to test items are usually summed to produce a posttest score for each student. These scores are, then, used to produce the average pretest and posttest scores for the course. A difference score is obtained by subtracting the pretest score from posttest score which represents the change or growth in knowledge and understanding over time (e.g., over the course of one semester).

However, using only a difference score is in some ways limiting, which is why a decomposition of the difference score as a composite was developed in a prior study (Walstad et al., in review). With the student's item responses across the two test measurements points four distinct patterns of learning were produced that could then be used to decompose the composite scores. These response patterns reflect four types of student learning, or at least they can be characterized as such for expository purposes of this study. The student learning perspective highlights the similarities and contrasts in the four outcomes and is useful for showing how each composite test score is influenced by the different types of learning. One key insight from this decomposition of test scores is that the difference score is constructed from two conflicting types of student learning, a result which raises further questions about its usefulness for measuring change (e.g., over a course of a study) under these test conditions. Criticism of the difference score, however, is not new: Decades ago Cronbach and Furby (1970, p. 80) recommended that "investigators who ask questions regarding gain scores would ordinarily be better advised to frame their questions in other ways". This study follows the spirit of this advice and presents an alternative approach to the difference score for assessing the gain in student learning. In the study presented in this paper, data from US-American and German students were analyzed to examine whether the observed effects are comparable.

One final point is worth stating as a pre-condition for understanding the test analysis. It is not known from the item responses why students select the particular pretest and posttest answers to items on a multiple choice test. There are many reasons such as subject understanding or guessing which vary by individual student (for findings from analyses of students' response processes and rationale see, e.g., Brückner & Pellegrino, 2016). The analyses presented in this paper offer an important basis for examining participants' reasons for selecting item responses.

By modeling four types of economic learning (see chapter 2), it is possible to gain in-depth insights into the data of the learning process while also taking guessing into consideration in the modelings and analyses. On the basis of the approach by Smith and Wagner (2017) and supplementary to the study by Walstad et al. (in review), the guessing parameter was included in the analysis of a pretest-posttest measurement using the Test of Understanding College Economics (TUCE; Walstad & Rebeck, 2007) as well as in the study by Happ et al. (2016) for a pretest-posttest measurement using the Test of Economic Literacy (TEL; Soper & Walstad, 1987). Modeling and controlling for the guessing parameter allows for much more valid statements on actual learning patterns (such as positive and negative learning) and its initial level. For example, if a participant achieved a high test score predominantly due to guessing, this cannot be considered positive learning. These kinds of results can offer valuable indications for further, more in-depth analyses at

the item level. The analyses presented here are important to gain more valid information about the acquisition of knowledge and understanding by students participating in a pretest-posttest-design, and they are important for both researchers and economics educators.

Conceptual Background

Response Patterns and Learning Scores

As mentioned before, four response patterns were introduced in the paper by Walstad et al. (in review). The first response pattern from matched pretest and posttest data would be supplying incorrect answers to test items on the pretest, but giving correct answers to the same items on the posttest. This pattern indicates that there is an improvement in student understanding, so an appropriate label for this outcome would be *positive learning* (PL). The second response pattern would be students giving correct answers to items on the pretest and the posttest. This outcome could be classified as *retained learning* (RL) because it appears to show that students maintained their understanding of content from pretest to posttest. These first two response patterns are generally considered as desired outcomes from a learning perspective because they show either an increase in understanding or at least the maintenance of understanding.

The remaining two response patterns provide information about what students apparently do not learn. For the third response set, students give correct answers to items on the pretest, but then deliver incorrect answers on the posttest. This response set indicates that there is a loss in understanding from pretest to posttest, so in contrast to previously described outcomes it can be characterized as *negative learning* (NL). The fourth response combination would be those items where a student gave wrong answers on both the pretest and posttest. It shows no change in student understanding, or what can be labeled as *zero learning* (ZL). Further ZL analysis could be conducted on the consistency of the incorrect responses, but it is a secondary matter, so the focus for this study will be on the general ZL condition of two incorrect responses.

From this learning perspective a difference score measures the total gain (incorrect to correct) (PL) after subtracting the total loss (correct to incorrect) (NL), or the "net" change in understanding. Whether it makes sense, however, to calculate the net change is questionable. The conceptual support for NL is less plausible than for PL because NL goes from understanding to not understanding, which suggests that guessing may have more of an influence on NL than PL. In addition, at the posttest, NL is the same as ZL because both responses are incorrect, but a difference score only uses NL responses. Subtracting NL from PL simply reduces the size of the total gain in understanding and the estimate of positive learning. By contrast, PL alone is less confounded by other factors and assumptions than a difference score, and thus can serve as a more direct measure of improvement in student understanding.

The above trade-offs are predicated on a restriction that holds constant the number incorrect. Dropping that restriction shows how a change in ZL affects the other learning scores. ZL is important because it influences the number incorrect on the pretest (ZL + PL) and the posttest (ZL + NL). If a "difficult" test is defined as one with many test items students cannot answer correctly on either the pretest or posttest, then ZL will be high. A high ZL, however, limits the size of PL and RL on the posttest and NL and RL on the pretest. The opposite is likely to be the case with a low ZL or "easy" test. The concepts discussed can be applied to economics test data to analyze its learning components.

Guessing

Taking into account that students guess if they do not know the right answer in a multiple choice test, a guessing parameter should be included in the modeling of learning and understanding. In cross-sectional designs, the discussion about and correction for guessing in multiple-choice tests has a long history (e.g., Lord, 1975; Frary, 1988; Espinosa & Gardeazabal, 2010). However, correcting for guessing in pretest-posttest designs has not been done very often in the past and according to the newly developed decomposition of the composite difference score in the four learning patterns, Smith and Wagner (2017) developed a suitable way to adjust for guessing for each learning and difference score within a pretest-posttest design.

Estimate	Expected Value	Limit ($n \to \infty$)
Post- Pre-test ($\hat{\text{flow}}$)	$(\mu - \alpha + \gamma + \frac{1}{n}(1 - \mu - \gamma + \alpha)) - (\mu + \frac{1}{n}(1 - \mu)) = \frac{(n-1)(\gamma-\alpha)}{n}$	$\gamma - \alpha$
Positive Learning ($\hat{\text{pl}}$)	$(1 - \gamma - \mu)\frac{n-1}{n}\frac{1}{n} + \gamma\frac{n-1}{n} = \frac{(n-1)(1-\mu+\gamma(n-1))}{n^2}$	γ
Negative Learning ($\hat{\text{nl}}$)	$\frac{n-1}{n}\alpha + \frac{1}{n}\frac{n-1}{n}(1 - \gamma - \mu) = \frac{(n-1)(1-\gamma-\mu+\alpha n)}{n^2}$	α
Retained Learning ($\hat{\text{rl}}$)	$\mu - \alpha + (1 - \gamma - \mu)\frac{1}{n}\frac{1}{n} + \gamma\frac{1}{n} + \alpha\frac{1}{n} = \mu - \alpha + \frac{1-\gamma-\mu}{n^2} + \frac{\gamma}{n} + \frac{\alpha}{n}$	$\mu - \alpha$
Zero Learning ($\hat{\text{zl}}$)	$(1 - \mu - \gamma)\frac{n-1}{n}\frac{n-1}{n} = \frac{(n-1)^2(1-\gamma-\mu)}{n^2}$	$1 - \gamma - \mu$

Table 1: Expected Value of Flow of Pre/Post Disaggregated Learning Types (Smith & Wagner, 2017, p. 7)

Based on the formulas for the expected values of the different learning scores (see Table 1), they introduced how to adjust for guessing by using a parameter n for the number of available answers in a task and the probability of guessing correct is 1/n . If n increases and the probability of guessing correctly decreases, positive learning equals the number of students who gave a wrong answer in the pretest and a correct answer in the posttest without guessing at all. This is what the authors define as γ. However, due to the fact that the number of possible answers isn't close to infinity, but in our case n=4 (with one correct answer and three distractors), we should adjust for guessing what leads to $\hat{\gamma}$, with $adj.pl = \hat{\gamma} = \frac{n(\widehat{nl}+\widehat{pl}n+\widehat{rl}-1)}{(n-1)^2}$ (Smith & Wagner, 2017, p. 8). The authors further define α as negative learning with the adjusted negative learning score $adj.nl = \hat{\alpha} = \frac{n(\widehat{nl}n+\widehat{pl}+\widehat{rl}-1)}{(n-1)^2}$ (ebd.). Retained learning is defined as μ and for adjusted retained learning it is $adj.rl = \hat{\mu} = \frac{\widehat{nl}+\widehat{rl}+1}{n-1} + \widehat{nl} + \widehat{rl}$ (ebd.). And finally, zero learning is the result of subtracting retained learning and positive learning from 1 and it is $adj.zl = 1 - \mu - \gamma$ (Smith & Wagner 2017, p. 5).

Following the formulas for positive, retained, negative and zero learning from Walstad and Wagner (2016) and Walstad et al. (in review) as well as the formulas for guessing adjustment (Smith & Wagner, 2017), we will analyze the four learning types with and without taking guessing into account in Section Results.

Design and Samples

Sample description of the German TEL study

In the first study we present here, a pretest-posttest design was administered with German students at the beginning of a degree course, before students had attended any learning opportunities in higher education economics (pretest). Over the course of one year, students attended introductory economics courses where basic economic concepts in microeconomics and macroeconomics were taught. After the first year, the economic knowledge and understanding of the German students was assessed a second time (posttest) (see Happ et al., 2016).

To measure economics knowledge and understanding, the German adaptation of the American *Test of Economic Literacy* (TEL) (Soper & Walstad, 1987) called "Wirtschaftskundlicher Bildungstest" (WBT) (Beck, Krumm, & Dubs, 2001) was used. The WBT has been proven to be a reliable and valid instrument for the assessment of knowledge and understanding in macro- and microeconomics. For the study presented in this paper, 14 TEL items were deemed relevant for higher education in Germany based on the content of introductory economics courses.

All participants of the German study (n=403) attended an introductory course in economics covering principles of economics as well as some basics of microeconomics and macroeconomics. We expected some degree of positive learning to occur from pretest to posttest for all participants. In accordance with Walstad and Wagner's model (2016), we examined whether and to what extent students exhibited other types of learning economics, as well and analyzed the influence of personal characteristics on their development of economic knowledge (see Happ et al., 2016).

Table 2: Sample description of the German TEL study (Happ et al., 2016)

	Sample (*N*)[a]	Mean pretest	Mean posttest
Total	403	7.47	9.14
Male	188	8.09	9.55
Female	215	6.93	8.77
German	364	7.54	9.17
Non-German	39	6.77	8.79
Vocational education	101	7.59	8.98
No vocational education	298	7.44	9.16
School major business and economics	123	7.85	9.06
Other major	280	7.31	9.17

Table 2 shows that in the sample of 403 German higher education students we surveyed 215 female students and 39 students with a non-German background. About one fourth of the surveyed students had completed vocational education prior to studies in higher education (101 persons) and almost one third graduated from a high school with business and economics major subjects. We can assume that due to this prior education, these students (30 percent of the sample) have a higher level of knowledge and understanding in economics, and therefore these 30 percent of the sample are less likely to resort to guessing when responding to the TEL items.

Sample description of the U.S. American TUCE study

The *Test of Understanding of College Economics* (TUCE) (Walstad, Watts, & Rebeck, 2007) is a nationally normed and standardized multiple choice test for principles of economics that was developed for use with undergraduate students in the United States. The TUCE consists of two tests, one for macroeconomics and one for microeconomics, which each contain 30 items. Data from the matched pretest and posttest sample who took the macro test (2,789 students in the United States) as well as who took the micro test (3,255 students in the United States) will be analyzed.

Table 3: Descriptive Statistics: Macro and Micro TUCE

	Macro sample			**Micro sample**		
Variables	N	Mean Pretest	Mean Posttest	N	Mean Pretest	Mean Posttest
Total	2,789	9.8	14.2	3,255	9.4	12.8
Male	1,651	10.2	14.8	1,848	9.6	13.1
Female	1,125	9.2	13.4	1,384	9	12.3
White	1,973	10	14.8	2,204	9.5	13-2
Non-White	804	9.3	12.8	1,015	9	11.8
Business major	1,453	9.5	13.9	1,637	9.1	12.2
Other major	1,309	10.1	14.5	1,549	9.7	13.4
English speaker	2,439	9.8	14.1	2,744	9.3	12.6
Non-English speaker	335	10	14.7	475	9.8	13.6
Took college economics courses	947	10.5	15.4	1,276	9.7	12.4
No college economic courses	1,835	9.5	13.6	1,958	9.2	13
Took high school economics courses	1,107	10	14.4	1,415	9.7	13
No high school economic courses	1,657	9.7	14	1,789	9.1	12.6

Results

To account for guessing according to the approach by Smith & Wagner (2017), the first step is the decomposition of learning patterns. This is why we estimated positive learning (PL), retained learning (RL), negative learning (NL), zero learning (ZL) and the difference between the pretest and posttest results (flow) for each item of the TUCE (see Tables 4 and 5) and for each item of the TEL (see Table 5). These estimates were then used to follow the guessing adjustment introduced by Smith and Wagner (2017) step by step to adjust the four learning scores as well as the flow.

Table 4: Guessing TUCE Macro (US Data)

	Unadjusted					Guessing-Adjustment					
	PL	RL	NL	ZL	Flow	μ	γ (adj. PL)	$\boldsymbol{\mu - \alpha}$ (adj. RL)	α (adj. NL)	$1 - \boldsymbol{\gamma - \mu}$ (adj. ZL)	$\gamma - \alpha$ (adj. flow)
q1	0.362	0.168	0.060	0.411	0.303	-0.030	0.300	0.073	-0.103	0.730	0.403
q2	0.259	0.348	0.146	0.247	0.114	0.325	0.236	0.241	0.084	0.439	0.152
q3	0.350	0.338	0.119	0.193	0.231	0.276	0.381	0.203	0.073	0.342	0.308
q4	0.239	0.219	0.143	0.399	0.096	0.150	0.142	0.135	0.014	0.709	0.128
q5	0.513	0.079	0.036	0.372	0.477	-0.180	0.519	-0.063	-0.118	0.662	0.636
q6	0.311	0.158	0.172	0.359	0.139	0.106	0.255	0.037	0.069	0.638	0.186
q7	0.267	0.333	0.179	0.221	0.088	0.349	0.258	0.209	0.140	0.393	0.118
q8	0.262	0.234	0.180	0.323	0.082	0.219	0.206	0.123	0.096	0.575	0.109
q9	0.238	0.088	0.131	0.544	0.107	-0.042	0.075	0.026	-0.068	0.967	0.143
q10	0.238	0.169	0.179	0.414	0.059	0.130	0.133	0.076	0.054	0.737	0.079
q11	0.332	0.261	0.080	0.327	0.252	0.121	0.297	0.161	-0.039	0.582	0.336
q12	0.313	0.239	0.156	0.293	0.157	0.193	0.287	0.115	0.078	0.520	0.209
q13	0.257	0.378	0.186	0.179	0.070	0.419	0.263	0.250	0.169	0.318	0.094
q14	0.336	0.139	0.110	0.415	0.226	-0.001	0.263	0.037	-0.038	0.738	0.301
q15	0.281	0.331	0.164	0.224	0.117	0.326	0.276	0.207	0.120	0.398	0.156
q16	0.269	0.112	0.166	0.454	0.103	0.036	0.157	0.017	0.019	0.807	0.138
q17	0.239	0.128	0.179	0.454	0.060	0.076	0.117	0.039	0.037	0.806	0.080
q18	0.361	0.087	0.087	0.465	0.275	-0.101	0.275	-0.011	-0.091	0.826	0.366
q19	0.249	0.147	0.179	0.425	0.070	0.101	0.143	0.052	0.050	0.755	0.094
q20	0.248	0.356	0.157	0.238	0.091	0.351	0.225	0.247	0.103	0.424	0.122
q21	0.299	0.119	0.085	0.497	0.214	-0.062	0.178	0.046	-0.108	0.884	0.286
q22	0.236	0.092	0.090	0.582	0.147	-0.091	0.056	0.048	-0.139	1.035	0.196
q23	0.256	0.107	0.157	0.479	0.099	0.019	0.129	0.023	-0.004	0.852	0.132
q24	0.225	0.103	0.159	0.513	0.066	0.016	0.072	0.032	-0.016	0.912	0.087
q25	0.314	0.286	0.151	0.249	0.163	0.249	0.308	0.159	0.090	0.443	0.218
q26	0.230	0.078	0.146	0.546	0.085	-0.035	0.064	0.014	-0.049	0.971	0.113
q27	0.249	0.076	0.097	0.578	0.152	-0.102	0.076	0.025	-0.127	1.027	0.203
q28	0.300	0.209	0.145	0.346	0.155	0.138	0.246	0.099	0.039	0.616	0.207
q29	0.226	0.115	0.132	0.526	0.094	-0.003	0.068	0.054	-0.058	0.936	0.125
q30	0.262	0.177	0.160	0.401	0.103	0.116	0.171	0.081	0.034	0.713	0.137
∅	**0.284**	**0.189**	**0.138**	**0.389**	**0.147**		**0.206**	**0.092**	**0.010**	**0.692**	**0.195**

Table 5: Guessing TUCE Micro (US data)

	Unadjusted					Guessing-Adjustment					
	PL	RL	NL	ZL	Flow	μ	γ (adj. PL)	$\boldsymbol{\mu} - \alpha$ (adj. RL)	α (adj. NL)	$1 - \boldsymbol{\gamma} - \mu$ (adj. ZL)	$\gamma - \alpha$ (adj. flow)
q1	0.257	0.244	0.143	0.355	0.114	0.183	0.185	0.150	0.033	0.631	0.152
q2	0.226	0.171	0.161	0.442	0.065	0.109	0.105	0.091	0.018	0.786	0.087
q3	0.292	0.210	0.146	0.353	0.146	0.141	0.232	0.103	0.037	0.627	0.195
q4	0.486	0.084	0.061	0.370	0.425	-0.141	0.483	-0.057	-0.084	0.658	0.567
q5	0.258	0.198	0.201	0.342	0.057	0.200	0.192	0.083	0.116	0.608	0.076
q6	0.307	0.149	0.083	0.461	0.224	-0.024	0.205	0.070	-0.094	0.819	0.299
q7	0.207	0.283	0.164	0.346	0.043	0.263	0.122	0.197	0.065	0.615	0.057
q8	0.226	0.142	0.065	0.567	0.161	-0.057	0.050	0.108	-0.165	1.007	0.215
q9	0.222	0.087	0.134	0.558	0.088	-0.040	0.048	0.030	-0.070	0.991	0.118
q10	0.233	0.206	0.160	0.401	0.073	0.155	0.132	0.120	0.035	0.713	0.097
q11	0.263	0.053	0.056	0.629	0.207	-0.189	0.071	0.016	-0.205	1.118	0.276
q12	0.333	0.118	0.124	0.425	0.209	-0.011	0.255	0.013	-0.023	0.755	0.279
q13	0.292	0.213	0.154	0.342	0.138	0.155	0.237	0.102	0.053	0.608	0.184
q14	0.310	0.143	0.162	0.385	0.148	0.073	0.242	0.029	0.045	0.684	0.197
q15	0.225	0.114	0.104	0.557	0.120	-0.042	0.052	0.066	-0.108	0.991	0.160
q16	0.251	0.245	0.182	0.321	0.069	0.237	0.192	0.137	0.100	0.571	0.092
q17	0.266	0.167	0.150	0.418	0.116	0.089	0.169	0.075	0.014	0.742	0.155
q18	0.261	0.150	0.149	0.441	0.112	0.065	0.152	0.063	0.002	0.783	0.149
q19	0.208	0.227	0.207	0.358	0.001	0.246	0.118	0.129	0.117	0.636	0.001
q20	0.247	0.060	0.110	0.583	0.137	-0.107	0.070	0.005	-0.112	1.037	0.182
q21	0.226	0.222	0.206	0.345	0.021	0.238	0.148	0.117	0.121	0.614	0.027
q22	0.237	0.355	0.200	0.208	0.037	0.406	0.223	0.232	0.174	0.370	0.049
q23	0.208	0.104	0.136	0.551	0.072	-0.012	0.032	0.051	-0.063	0.980	0.095
q24	0.268	0.221	0.184	0.327	0.084	0.207	0.212	0.106	0.100	0.582	0.111
q25	0.219	0.116	0.118	0.547	0.101	-0.021	0.048	0.065	-0.086	0.973	0.135
q26	0.229	0.113	0.181	0.476	0.049	0.059	0.094	0.030	0.029	0.847	0.065
q27	0.258	0.151	0.141	0.451	0.117	0.055	0.144	0.068	-0.013	0.801	0.156
q28	0.242	0.107	0.137	0.514	0.105	-0.008	0.094	0.038	-0.046	0.914	0.140
q29	0.226	0.145	0.163	0.466	0.064	0.077	0.095	0.067	0.010	0.829	0.085
q30	0.259	0.234	0.171	0.335	0.088	0.207	0.197	0.128	0.080	0.596	0.117
∅	**0.258**	**0.168**	**0.145**	**0.429**	**0.113**		**0.153**	**0.081**	**0.003**	**0.763**	**0.151**

With a view to the TUCE results (see Tables 4 and 5), it is evident that on average, PL and RL but also NL is expected to be obviously overestimated, whereas ZL is apparently underestimated (0.389 vs. the adjusted score of 0.692 for macro and 0.429 vs. the adjusted score of 0.763 for micro). Between macro and micro the underestimation of ZL seems to be of the same amount, but the overestimation for PL is a bit higher in the micro items. Remarkably, when taking into consideration the guessing adjustment, there is almost no NL (0.01 in the macro items and 0.03 in the micro items).

Table 6: Guessing TEL (German data)

	Unadjusted					Guessing-Adjustment					
	PL	**NL**	**RL**	**ZL**	Flow	μ	γ (adj. PL)	α (adj. NL)	$\mu - \alpha$ (adj. RL	$1 - \gamma - \mu$ (adj. ZL)	$\gamma - \alpha$ (adj. flow)
C1	0.31	0.08	0.51	0.09	0.23	0.46	0.37	0.07	0.39	0.17	0.30
C2	0.36	0.09	0.37	0.17	0.27	0.29	0.41	0.04	0.24	0.31	0.36
C5	0.13	0.10	0.62	0.14	0.03	0.63	0.11	0.07	0.55	0.26	0.04
C6	0.14	0.06	0.03	0.78	0.08	-0.22	-0.16	-0.27	0.05	1.39	0.11
C7	0.22	0.04	0.72	0.02	0.18	0.68	0.28	0.04	0.64	0.04	0.24
C8	0.23	0.09	0.62	0.06	0.14	0.61	0.28	0.09	0.52	0.11	0.19
C9	0.25	0.17	0.37	0.21	0.08	0.38	0.24	0.13	0.25	0.38	0.11
C10	0.21	0.16	0.42	0.21	0.05	0.44	0.18	0.12	0.33	0.37	0.06
C11	0.17	0.08	0.69	0.06	0.09	0.69	0.21	0.08	0.61	0.11	0.13
C12	0.20	0.11	0.40	0.30	0.09	0.34	0.13	0.01	0.33	0.53	0.12
C13	0.17	0.15	0.48	0.20	0.02	0.51	0.14	0.11	0.40	0.35	0.03
C14	0.32	0.09	0.23	0.36	0.23	0.09	0.27	-0.04	0.14	0.64	0.31
C15	0.25	0.16	0.34	0.24	0.09	0.34	0.23	0.11	0.23	0.43	0.12
C17	0.23	0.16	0.13	0.48	0.07	0.05	0.09	-0.01	0.06	0.86	0.09
∅	**0.228**	**0.109**	**0.425**	**0.239**	**0.119**		**0.198**	**0.039**	**0.339**	**0.424**	**0.159**

The TEL data set (Table 6) largely replicates the results of the analysis using the TUCE data set (see Tables 4 and 5). This is interesting, as the two samples were collected in different countries. The slightly smaller effects could also be traced back to the much smaller sample size (N=403 vs. N=2,789 for TUCE macro and 3,255 for TUCE micro). However, taking guessing into account leads to a higher proportion of zero learning, i.e., a learning pattern where students did not know or correctly guess solutions, neither in the pretest nor in the posttest. For the TEL, we also found that PL, RL, NL were slightly overestimated, but not as much as for the TUCE data. Accordingly, ZL was slightly underestimated for the TEL pretest-posttest measurements, but with 0.239 vs. 0.424 after adjusting for guessing, the difference is not as high as it was for the TUCE.

Overall, as expected, the analyses for both data sets indicate that positive, negative and retained learning were slightly *over*estimated when not taking guessing into account. Taking guessing into account, the proportion of zero learning is much higher, which means that without guessing, ZL is, as expected, obviously *under*rated.

Discussion

With a view to our claim made in the introduction, the results show that without taking guessing into account, the two learning patterns PL and ZL in particular are at times severely skewed, as a remarkable proportion of PL can be explained by guessing. Vice versa, when including the guessing parameter, the adjustment is higher for NL than for PL. However, for a valid interpretation of the results, more in-depth analyses at the cognitive level as well as the content level are necessary, in order to gain further indications about learning patterns. One possibility for further modeling is the 'thresholds' approach, which makes it possible to model the transition of students' economics understanding from a subject-related basic threshold to higher levels of understanding (for a modeling example using TUCE IV data, see Brückner & Zlatkin-Troitschanskaia, 2018).

References

Beck, K., V. Krumm, & R. Dubs (2001). WBT—Wirtschaftskundlicher Bildungstest [German adaptation of the Test of Economic Literacy (TEL)]. In Handbuch wirtschaftspsychologischer Testverfahren [Guide to assessment methods of business psychology], ed. W. Sarges and H. Wottawa, 559–62. Lengerich, Germany: Pabst Science Publishers.

Brückner, S., & Pellegrino, J. W. (2016). Integrating the analysis of mental operations into multilevel models to validate an assessment for higher education students' competency in business and economics. *Journal of Educational Measurement, 53*(3), 293-312.

Brückner, S., & Zlatkin-Troitschanskaia, O. (2018). Threshold concepts for modeling higher education students' understanding and learning in economics - implications for instruction and assessment. In O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper, C. Lautenbach, & C. Kuhn (Eds.). *Assessment of Learning Outcomes in Higher Education – Cross National Comparisons and Perspectives.* Wiesbaden: Springer.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change"—or should we? *Psychological Bulletin, 74*(1), 68–80.

Espinosa, M. P., & Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology*, 54(5), 415-425.

Frary, R. B. (1988). Formula scoring of multiple-choice tests (correction for guessing). *Educational Measurement: Issues and Practice*, 7(2), 33-38.

Happ, R., Zlatkin-Troitschanskaia, O., & Schmidt, S. (2016). An Analysis of Economic Learning among Undergraduates in Introductory Economics Courses in Germany. *The Journal of Economic Education*, 47(4), 300–310. doi:10.1080/00220485.2016.1213686

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement*, 12(1), 7-11.

Smith, B. O. & Wagner, J. (2017). *Adjusting for Guessing and Applying a Statistical Test to the Disaggregation of Value-Added Learning Scores* (April 20, 2017). Available at SSRN: https://ssrn.com/abstract=2941454

Soper, J., & W. Walstad (1987). *Test of economic literacy: Examiner's manual*. 2nd ed. New York: Council for 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.

Walstad, W. B., Rebeck, K., &, Butters, R. B. (2013). *Test of Economic Literacy* (4th edition): Examiner's Manual. New York: Council for Economic Education.

Walstad, W. B. & Wagner, J. (2016). The disaggregation of value-added test scores to assess learning outcomes in economics courses. *The Journal of Economic Education*, 47(4), 121-131.

Walstad, W. B., Schmidt, S. & Zlatkin-Troitschanskaia, O. (in review). *Measuring Change in Economic Understanding*.