A Comparison of Different Item Response Theory Models for Scaling Speeded C-Tests

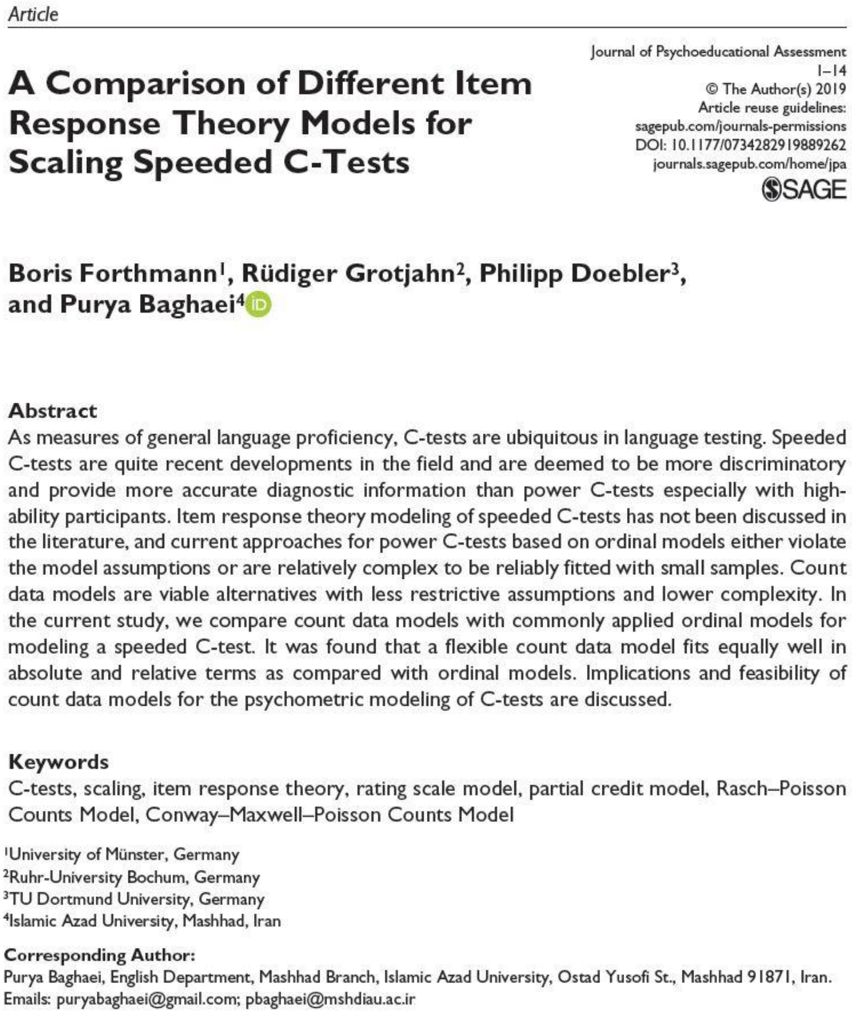
Article *in* Journal of Psychoeducational Assessment · December 2019

DOI: 10.1177/0734282919889262

Boris Forthmann
University of Münster
94 PUBLICATIONS **1,636** CITATIONS

Philipp Doebler
Technische Universität Dortmund
100 PUBLICATIONS **1,916** CITATIONS

Purya Baghaei
Islamic Azad University Mashhad Branch
92 PUBLICATIONS **1,402** CITATIONS



Article

Journal of Psychoeducational Assessment
1–14
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0734282919889262
journals.sagepub.com/home/jpa
SAGE

A Comparison of Different Item Response Theory Models for Scaling Speeded C-Tests

Boris Forthmann[1], Rüdiger Grotjahn[2], Philipp Doebler[3], and Purya Baghaei[4]

Abstract

As measures of general language proficiency, C-tests are ubiquitous in language testing. Speeded C-tests are quite recent developments in the field and are deemed to be more discriminatory and provide more accurate diagnostic information than power C-tests especially with high-ability participants. Item response theory modeling of speeded C-tests has not been discussed in the literature, and current approaches for power C-tests based on ordinal models either violate the model assumptions or are relatively complex to be reliably fitted with small samples. Count data models are viable alternatives with less restrictive assumptions and lower complexity. In the current study, we compare count data models with commonly applied ordinal models for modeling a speeded C-test. It was found that a flexible count data model fits equally well in absolute and relative terms as compared with ordinal models. Implications and feasibility of count data models for the psychometric modeling of C-tests are discussed.

Keywords

C-tests, scaling, item response theory, rating scale model, partial credit model, Rasch–Poisson Counts Model, Conway–Maxwell–Poisson Counts Model

[1]University of Münster, Germany
[2]Ruhr-University Bochum, Germany
[3]TU Dortmund University, Germany
[4]Islamic Azad University, Mashhad, Iran

Corresponding Author:
Purya Baghaei, English Department, Mashhad Branch, Islamic Azad University, Ostad Yusofi St., Mashhad 91871, Iran.
Emails: puryabaghaei@gmail.com; pbaghaei@mshdiau.ac.ir

C-tests are general measures of first or second language proficiency (Raatz & Klein-Braley, 1981) and have been successfully developed for various languages (Norris, 2018). C-tests are widely used in research (for example, in studies on language proficiency of bilinguals; Hulstijn, 2012) and educational practice (Grotjahn, 2019) such as placement (Eckes, 2011; Mozgalina & Ryshina-Pankova, 2015), or remedial teaching (Linnemann & Wilbert, 2010).

C-tests were suggested as improvements over cloze tests (Klein-Braley, 1997). While in a cloze test every *n*th word is deleted in a single passage a C-test consists of several passages in which starting from the second word of the second sentence the second half of every second or sometimes every third word is deleted (Grotjahn, 2016). Subjects have to restore the deleted parts and the score on a passage is based on the number of gaps correctly reconstructed. A C-test battery usually contains four to six passages, each with 25 or 20 gaps. C-tests are mostly conceived of as more or less pure power tests (Fadaeipour & Zohoorian, 2017), and, therefore, rather generous time limits are set for the whole test.

A C-test in which the time limit of the individual passages is fixed during administration, such that test-takers have to work under time pressure has been referred to by Grotjahn (2010) as a speeded C-test (S-C-test). S-C-tests assess the ability to quickly process language under time pressure, which is considered an important component of the construct of language ability (Alderson, 2005). Moreover, highly speeded S-C-tests (1 to 2 minutes per passage) allow assessing ability differences even at the highest proficiency levels (Grotjahn, 2010; Wockenfuß & Raatz, 2014).

Item Response Theory Modeling of C-tests

In principle, each gap in a C-test could be considered as an item. However, this results in a massive increase in the number of item parameters to be estimated and, as the gaps are nested in texts, conditional independence could be violated. Conditional independence is a basic assumption of all IRT and Rasch models and the violation of this assumption results in biased parameter estimates and spuriously high reliability coefficients (Sireci, Thissen, & Wainer, 1991). In addition, when C-test texts are presented under speeded conditions examinees usually will not reach the gaps towards the end of the passages which makes treating gaps as individual items problematic (Heckman, Tiffin, & Snow, 1967).

Given these problems, researchers have most often employed the item bundle approach (Rosenbaum, 1988). In this modeling strategy, the scores on the gaps are aggregated as passage scores (testlets or super items) and then entered into Rasch model analysis (Eckes & Baghaei, 2015). That is, each passage is considered as a polytomous item with 20-25 ordered categories and an ordinal Rasch model like the rating scale model (RSM; Andrich, 1978) or the partial credit model (PCM; Masters, 1982) is applied.

One drawback of these models for scaling C-tests is that they have many parameters to be estimated and thus require large sample sizes which are not always available. According to Eckes (2011), 10 observations per response category (i.e., the number of solved gaps) are required for stable estimation of PCM parameters. This can easily result in sample size requirements of 500 or more depending on the number of gaps in each passage, the number of passages, and the mapping of person proficiency and item difficulty (see Eckes, 2011). For example, with six C-test items with 25 gaps a minimum sample size of 260 would be required, but only when assuming that participants are uniformly distributed across all possible passage scores ranging from 0 to 25 for each of the six items. However, no examinee is likely to fill all the gaps in a passage within the allotted time in S-C-tests. This leaves many high passage scores unobserved or with very few observations and, as a consequence, leads to biased person parameter estimates and misleading estimation accuracy (Li, 2013).

Furthermore, the RSM assumes similar rating scale structure and equal distances between the categories of the rating scale for all the items. This is a very strong and unrealistic assumption regarding S-C-tests especially when unequal time limits are set for each passage. Finally, item difficulty parameters in RSM and PCM are only general item main effects and the time limits for the items are not taken into consideration. If the time limit for an easy text is short and its total raw score is smaller than the total raw score for a difficult text with a longer time limit, the easier texts will turn out to have a higher difficulty parameter. In essence, applying RSM and PCM to rather small sample sizes is problematic. Furthermore, these models do not allow disentangling the effect of time limits from item difficulty.

A flexible count data approach for speeded C-tests. In addition to the Rasch model for dichotomous data, Rasch (1960) proposed a model for count data known as the Rasch Poisson Counts Model (RPCM; e.g., Baghaei & Doebler, 2018). One restriction of this model is that conditional on person ability and item parameter, mean and variance are equal. This strong model assumption of equidispersion makes the model less useful than other members of the Rasch model family. While subsequent methodological advances somewhat widened the applicability of the RPCM (Hung, 2012), only recently Authors (2019) showed how the original RPCM can be generalized into the Conway Maxwell Poisson Counts Model (CMPCM). The CMPCM is of similar flexibility as many linear latent trait models, as the mean is not directly related to the error variance. In the RPCM and CMPCM the expected value μ_{ji} of the score of person *j* on item *i* is modeled as a function of person ability θ_j and item easiness parameter β_i (i.e., larger values leading to higher scores on average):

$$\mu_{ji} = \exp(\beta_i + \theta_j).$$

The CMPCM generalizes the RPCM by including item-specific dispersion parameters ν_i, which determine the variance of the item residuals. It is convenient to report item dispersion on a log-scale, i.e., we use $\tau_i = \ln(1/\nu_i)$. This has the advantage that values of $\tau_i > 0$ indicate more variance than the Poisson distribution (overdispersion) and $\tau_i < 0$ stand for less variance (underdispersion). An even more parsimonious model is obtained when all dispersion parameters are set equal, $\tau_i = \tau$ for all *i*, which is called the CMPCM with global dispersion. Furthermore, in the CMPCM, the time limits can be included as an offset for the item easiness. Hence item easiness parameters can be interpreted as average scores per time unit (e.g., minutes): The expected number of hits or errors made by a person of average ability per unit of time is given by the easiness parameter. If the τ_i are fixed to zero the RPCM is obtained, hence the CMPCM generalizes the RPCM. The person parameter θ_j is assumed to be a Gaussian latent variable, with variance σ_θ^2 estimated freely. Moreover, the CMPCM, as compared to RSM and PCM, is a model with fewer parameters (two parameters per item and a latent ability variance) and can be estimated with small sample sizes (Authors, 2019).

The Present Research

C-tests are commonly modeled with ordinal item response theory models, which has several drawbacks as discussed above. Therefore, in the present study count data models are used for the scaling of a speeded C-test (S-C-test) and the results are compared with ordinal models. In addition, based on the count data models, an informative multiple group modeling approach analogous to measurement invariance (MI) testing in linear models (e.g., Vandenberg & Lance, 2000) is described, which takes a priori known differences in language proficiency into account (i.e., native speakers vs. second language learners). Generally speaking, the objective is to test whether the latent scales in both proficiency groups are one and the same, so that a latent ability of, say, 0.2 has the same interpretation regardless of group membership.

Towards this purpose, a series of model comparisons is performed within the CMPCM framework. First, a test is needed to decide whether the RPCM, CMPCM with global dispersion, or the CMPCM with item-specific dispersion fits best in the groups. Then, the chosen model needs to be estimated as a multiple group model with group-specific item easiness, item dispersion, latent variable variance, mimicking the configural invariance model (Vandenberg & Lance, 2000). Then, a strong invariance model can be formulated by constraining item easiness parameters to be equal across groups and estimating the latent mean difference. If this model holds, the latent scales of both groups are the same. Next, the strong invariance model can be further restricted by also constraining the dispersion parameters to be equal across groups which implies strict invariance (i.e., equal error variance and the same local reliability). These models are then compared with likelihood ratio tests.[1]To address the issues above the following research questions were formulated:

1. Does the CMPCM fit the data?
2. Which count data model fits the data best (RPCM, CMPCM with global dispersion or CMPCM with item-specific dispersion)?
3. Do the count data models fit better than the RSM and PCM?
4. To what extent are person parameter estimates from the count data models and ordinal IRT models correlated?

Method

Participants

The data were collected within two research projects and consist of 271 participants (Grotjahn, Schlak, & Aguado, 2010; Heine, 2017). Participants with missing data for age of onset (*n* = 14) had to be discarded from analysis.[2] This resulted in a sample of 257 participants from four different populations (see Heine, 2017): a) monolingual German natives, b) early bilinguals (onset age ≤ 3), c) late starters (onset age ≥ 16), and d) young starters (4 ≤ onset age < 16). Monolinguals and early bilinguals formed a native speaker group and late and young starters formed a second language learner group for analyses. In addition, we identified one multivariate outlier by means of the Mahalanobis distance as suggested by Tabachnick and Fidell (2007). This participant was discarded from analyses. Sample sizes and some pertinent characteristics of the participants are shown in Table 1.

Instrument

The S-C-test analyzed in this study was originally developed by Grotjahn et al. (2010) in the context of a research project on the effects of age on ultimate attainment in second language acquisition and was intended to differentiate among extremely competent second language learners of German. The S-C-test contained six passages, each containing 25 blanks, with text specific time limits ranging from 65 to 115 seconds. A more detailed description of the process of test development and the rationale for setting the time limits can be found in Grotjahn et al. (2010). Grotjahn et al. (2010) and Heine (2017) also reported excellent reliability estimates from .96 (*n* = 37) to .99 (*n* = 269). Moreover, Grotjahn et al. (2010) found a correlation of .63 (*p* < .001) with the subjects' self-evaluation of their German language competence, and Heine (2017) reported positive correlations between the present S-C-test and motivation to learn German as a foreign language, usage of German in everyday life, and prior participation in German language lessons (*r*s ranged from .26 to .33; all *p*s < .019).

Analytic Strategy

Ordinal models (RSM and PCM) were estimated with the R package mirt (Chalmers, 2012). We employed mirt because it is one of the few packages that allow estimating ability parameters by means of the maximum a posteriori (MAP) estimator (e.g., De Boeck et al., 2011). This estimator was required for the purpose of comparison because glmmTMB (Brooks et al., 2017), the package for CMPCM estimation, also uses MAP for ability parameter estimation. Using the same ability estimator for all models ensured that comparing models in terms of absolute fit by means of covariate-adjusted frequency plots (CAFPs; Holling, Böhning, Böhning, & Forman, 2016), correlations between ability estimates, and empirical reliability were not affected by different choices of ability estimators. Here, a CAFP indicates good fit, if the model implied total score distribution is close to its empirical counterpart. In contrast to fit indices, CAFPs do not boil the fit down to one number, but deviations from fit can be detected for each potential score. The data and analysis script are available in the Open Science Framework: https://osf.io/gbtd3/?view_only=8acb0468c24949dba3a63a9b953ae204

Results

Comparison of Count Data Models with Ordinal Models

First, the RSM, PCM, RPCM, and CMPCMs were compared in terms of absolute fit by means of a CAFP. Figure 1 shows that the more flexible ordinal models fit better towards the upper tail close to the maximum possible score of 25. Apparently, the data suffer from a slight ceiling effect in the group of native speakers and the ordinal models were better in accounting for this. Figure 1 further illustrates that the RPCM with its assumption of equidispersion was too inflexible to account for the observed counts. The CMPCMs provided much better fit to the data as compared to the RPCM (further supported by a likelihood ratio test and information criteria results; see Table 4). Another slight improvement of model fit at the upper tail was observed when dispersion was modeled at the item level as compared to the CMPCM with global dispersion (also further supported by likelihood ratio test and AIC results; see Table 4), while the BIC prefers the simpler CMPCM with global dispersion. One way to interpret the discrepancy of AIC and BIC is that the model with global dispersion might have greater explanatory power, while the item level dispersion CMPCM could provide better predictions (Vrieze, 2012). As their mean structures (Table 4) and the resulting reliability estimates (see below) are virtually identical, the discrepancy might be insubstantial. The fit of the CMPCMs was much closer to the fit of the ordinal variables and in light of fewer parameters in the CMPCMs, we conclude that these models displayed decent fit to the data in comparison with the RSM and PCM.

The parameter estimates of the count data models are depicted in Table 4 (item descriptive statistics are shown in Table 3). The item easiness parameters were comparable across RPCM and the CMPCMs. Importantly, the easiness parameters are interpretable independent of the allotted time to work on a text (the item time limits are shown in Table 3). Item 2 appeared to be the easiest item, whereas Item 6 was estimated to be hardest. In addition, the latent ability variance was slightly higher for the CMPCMs as compared to the RPCM. Both CMPCMs highlighted the presence of underdispersion in the data. When dispersion was modeled at the item level, Item 2 exhibited the highest degree of underdispersion, whereas Item 6 was closest to a value of zero (but still displaying a moderate level of underdispersion).

Subsequently, ability estimates from the RSM, PCM, RPCM, and CMPCMs were correlated to assess whether the order of participants was comparable across models (see Table 2). All correlations between ability estimates were > .95 which highlights that from a practical point of view the same latent variable was operationalized regardless of model choice (Larwin & Milton, 2012). In addition, empirical reliability estimates were quite comparable across the models and generally excellent (all $Rel(\theta) > .942$). In particular, reliability estimates were the same (when rounded to three decimals) for both ordinal models ($Rel(\theta) = .961$) and the CMPCM models ($Rel(\theta) = .969$) clearly performed on par with these estimates. The slightly lower reliability of the RPCM as compared to the CMPCM models was due to the underdispersion in the data that is not taken into account by the RPCM (Authors, 2019).

Moreover, we compared how the person specific standard errors of the ability estimators behaved, i.e., we compared conditional reliability estimates. Conditional reliabilities are depicted in Figure 2 as a function of the models and ranked ability estimates (to ensure comparability). Notably, the standard errors (*SE*s) differ between ordinal and count data models. That is, for ordinal models the *SE* depends on the distance of average test easiness parameters and person ability (i.e. *SE* is lowest when ability and test easiness match). Hence, persons with very low or very high ability have lower conditional reliability when the average test easiness matches average ability. In such situations an inverted U-shape for conditional reliability as a function of ability is implied and is clearly indicated for RSM and PCM in Figure 2. Otherwise, empirical reliability with the count data models was found to be increasing with the ability estimates which was expected for the RPCM (Graßhoff, Holling, & Schwabe, 2018), but similarly was found for the CMPCMs. Moreover, CMPCMs can model underdispersion in the data. Thus, these models were more informative and yielded better reliability estimates as compared to the RPCM.

With the RPCM as an exception, conditional reliability for the lowest ability estimates was found to be very close to .90 for all other models which implies excellent reliability (see Figure 2). Only from the lowest ranks towards the middle of the ability ranks, conditional reliability was slightly higher for the ordinal models as compared to the CMPCMs. From the middle of the ability ranks to the highest level, conditional reliability increased further and was slightly higher for the CMPCMs as compared to the ordinal models, whereas the ordinal models dropped drastically from rank 200 onwards (see Figure 2). The latter observation further implied the largest differences in terms of conditional reliability between the CMPCMs and the ordinal models at the upper tail of the ability distributions.

MI Testing for Native Speakers vs. Second Language Learners

First, a configural model was chosen in both groups separately. In both groups the CMPCM with global dispersion fitted better to the data as compared to the RPCM (native speakers: $\Delta\chi^2(1) = 209.199$, $p < .001$; second language learners: $\Delta\chi^2(1) = 53.528$, $p < .001$). Next, the CMPCM with item-specific dispersion fitted better to the data as compared to the CMPCM with global dispersion in both groups (native speakers: $\Delta\chi^2(5) = 34.770$, $p < .001$; second language learners: $\Delta\chi^2(5) = 19.917$, $p = .001$). These results were further supported by group-specific CAFPs which are shown in Figure 3. Analogous to $\Delta\chi^2$ testing, the differences between the models were more pronounced in the native speaker group (right plot in Figure 3) as compared to the second language learner group (left plot in Figure 3). It is evident from Figure 3 that all count data models fit much better to the data of the second language learner group as compared to the native speaker group in which a slight ceiling effect is likely to deteriorate model fit. The best fit of the CMPCM with item-specific dispersion in both groups was driven by underdispersion (range of estimates across both groups: -1.81 to -0.08).

Then, the CMPCM with item-specific dispersion was fit with group-specific item easiness, item dispersion, latent variable variance, and the latent mean unmodeled. This model was compared to a strong invariance model in which item-easiness parameters were constrained to be equal across groups and the latent mean difference was estimated. The strong invariance model deteriorated model fit as compared to the configural invariance model ($\Delta\chi^2(5) = 22.932$, $p < .001$). A more thorough investigation of item×group interaction revealed that two items differed (Item 4 and Item 6; see Figure 5) in terms of their deviation from unweighted average item easiness (i.e., effect coding for items was used) between native speakers even when *p*-values were adjusted for multiple testing (see Figure 5). Thus, we concluded that configural invariance is defensible, but no stricter levels of MI could be achieved (however, for illustration purposes a dispersion invariant model was also fitted and depicted in Figure 4).

The configural invariant model included effect coded item parameters (i.e., the intercept represents the unweighted average easiness across items and item coefficients represent the difference to the intercept). The item×group interaction was also included and, thus, an overall difference in average item easiness between native speakers and second language learners and also differences in item coefficients between groups were estimated. The overall estimate indicated that the number of gaps solved for a one-minute text would reduce by a factor of 0.485 (95%-CI: [0.441, 0.534]) for the second language learner group as compared to the native speaker group (i.e., roughly the expected number correct per text should be halved for second language learners as compared to native speakers).

Discussion

Psychometric modeling of C-tests in general is often based on ordinal IRT models and, thereby, the application of overly complex models and violations of model assumptions are generally tolerated. In the present study, we suggested a shift from common practice in C-test psychometric modeling towards flexible count data alternatives. Hence, we have applied these models to S-C-tests which are exemplary of C-tests and are further advantageous in terms of construct coverage and the assessment of individual differences at higher proficiency levels as compared to untimed C-tests. It was found that the CMPCM with item-specific or global dispersion is flexible enough to compete with RSM and PCM. The CMPCM fits comparably well in terms of absolute fit, ordering of persons' ability estimates, and average measurement precision.

We have further illustrated the feasibility of more complex analyses within the CMPCM framework by testing MI of a German S-C-test across groups of native speakers and second language learners. Given that C-tests are considered to be measures of general language proficiency that can be applied to diverse groups (e.g., monolinguals, early bilinguals, late or early second language learners), a test of MI is essential to establish psychometric quality (Baghaei, Bensch, & Ziegler, 2016). In this vein, we found that S-C-tests satisfy configural MI which is essential for construct validity. The findings are partially in line with previous research on MI of C-Tests. Baghaei (2010) examined the stability of Rasch model item parameters across low and high ability examinees. Findings showed that item parameters were relatively stable. Reichert, Brunner, and Martin (2014) examined the MI of a French and a German C-test across students speaking German/ Luxembourgish and those speaking Romance languages. Multi-group confirmatory factor analysis established configural invariance. However, as it was the case in the current study, metric and scalar invariance could not be supported. They issue some warning on the use of C-tests in high stakes testing.

Indeed, invariant C-tests across groups are essential when we want to compare them with respect to language proficiency without confounding proficiency differences by group-dependent properties (e.g., differential item functioning). Since RSM and PCM (the latter in particular) have a lot more parameters as compared to the CMPCM (i.e., parameters are essentially estimated twice in MI testing). As a consequence, standard errors of many model parameters are potentially smaller. Studies planning to use CMPCMs, for the purpose of MI testing, can hence plan with lower sample sizes. Thus, the CMPCM is expected to aid the development of invariant (S-)C-tests, but we caution that statistical planning should be backed up by simulations..

Limitations

In the present study, we examined the fit of a German C-test to CMPCM. Future studies should examine the applicability of count data IRT models to C-tests in other languages.

In addition, we observed a ceiling effect for the sample of native speakers. Thus, discriminating participants at the highest level of language proficiency was not possible. The observations capped at the maximum passage scores are, statistically speaking, right-censored. If the passages had been longer and contained further blanks, the ceiling effects could have been avoided. Further methodological research is needed, so that right-censoring can be accounted for by the CMPCM.

In the present study, we only demonstrated configural invariance for the S-C-test. However, our approach to invariance testing has been strict in the sense that only chi-squared difference tests have been studied, in contrast to established methods for structural equation models a significant test is informed by also calculating the difference in comparative fit index (ΔCFI; Cheung & Rensvold, 2002). However, further research will have to propose an analogous effect size for count data IRT models, so that strong invariance is not to be entirely discarded.

Conclusion

Item response theory is commonly used for the analysis and scaling of large-scale tests including the Program for International Student Assessment (PISA) and the National Assessment of Educational Progress (OECD, 2017). It is also used for modelling the TOEFL (Test of English as a Foreign Language; see www.ets.org/toefl) and the Testdaf (Test of German as a Foreign Language; see www.testdaf.de). Furthermore, IRT is frequently used to provide validity evidence for new psychoeducational tests (Baghaei & Tabatabaee-Yazdi, 2016).The current study demonstrated that the CMPCM provides a viable approach for the scaling of S-C-tests (and most likely C-tests in general).

Field Code Changed

Formatted: English (United States)

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. New York, NY: Continuum.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573. https://doi.org/10.1007/BF02293814

Baghaei, P., & Grotjahn, R. (2014). Establishing the construct validity of conversational C-tests using a multidimensional Rasch model. *Psychological Test and Assessment Modeling*, *56*, 60-82.

Baghaei, P., Bensch, D., & Ziegler, M. (2016). Measurement invariance across gender and major in the University of Tehran English Proficiency Test. In Aryadoust, V., & Fox, J. (Eds.), *Trends in Language Assessment Research and Practice: The View from the Middle East and the Pacific Rim* (pp.167-183). New Castle, England: Cambridge Scholars.

Baghaei, P., & Tabatabaee-Yazdi, M. (2016). The logic of latent variable analysis as validity evidence in psychological measurement. *The Open Psychology Journal*, *9*, 168-175. https://doi.org/10.2174/1874350101609010168

Baghaei, P. & Doebler, P. (2018). Introduction to the Rasch Poisson counts model: An R tutorial. *Psychological Reports.* Advance online publication. https://doi.org/10.1177/0033294118797577.

Baghaei, P. (2010). An investigation of the invariance of Rasch item and person measures in a C-Test. In R. Grotjahn (Ed.). *Der C-Test: Beiträge aus der aktuellen Forschung/ The C-Test: Contributions from Current Research* (pp.100-112). Frankfurt/M.: Lang.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., … Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*, 378–400. doi:10.32614/RJ-2017-066

Brown, A. & Croudace, T. J. (2015). Scoring and estimating score precision using multidimensional IRT. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Application to typical performance assessment (a volume in the multivariate applications series)*. New York, NY: Routledge/Taylor & Francis Group.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233-255. https://doi.org/10.1207/S15328007SEM0902_5

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*, 1-28. https://doi.org/10.18637/jss.v039.i12

Eckes, T. (2011). Item banking for C-tests: A polytomous Rasch modeling approach. *Psychological Test and Assessment Modeling*, *53*, 414-439.

Field Code Changed

Formatted: English (United States)

Eckes, T., & Baghaei, P. (2015). Using testlet response theory to examine local dependence in C-tests. *Applied Measurement in Education*, *28*, 85-98. https://doi.org/10.1080/08957347.2014.1002919

Fadaeipour, A., & Zohoorian, Z. (2017). Comparing psychometric characteristics of speed and standard C-test. *International Journal of Language Testing*, *7*, 40-50.

Graßhoff, U., Holling, H., & Schwabe, R. (2018). D-optimal design for the Rasch counts model with multiple binary predictors. *arXiv preprint arXiv:1810.03893*.

Grotjahn, R. (2010). Gesamtdarbietung, Einzeltextdarbietung, Zeitbegrenzung und Zeitdruck: Auswirkungen auf Item- und Testkennwerte und C-Test-Konstrukt. In R. Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from current research* (pp. 265-296). Frankfurt, Germany: Lang

Grotjahn, R. (2016, February 9). The electronic C-Test bibliography: Version 2016. Retrieved from http://www.c-test.de

Grotjahn, R. (2019). C-Tests. In S. Jeuk & J. Settinieri (Eds.), *Sprachdiagnostik Deutsch als Zweitsprache: ein Handbuch* (pp. 579-603). Berlin, Germany: De Gruyter Mouton.

Grotjahn, R., Schlak, T., Aguado, K. (2010). S-C-Tests: Messung automatisierter sprachlicher Kompetenzen anhand von C-Tests mit massiver textspezifischer Zeitlimitierung. In R. Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from current research* (pp. 297-319). Frankfurt, Germany: Lang.

Heckman, R. W., Tiffin, J., & Snow, R. E. (1967). Effects of controlling item exposure in achievement testing. *Educational and Psychological Measurement*, *27*, 113-125. https://doi.org/10.1177/001316446702700111

Heine, S. (2017). *Fremd- und Zweitsprachenerfolg und seine Erklärung durch Erwerbsalter, kognitive, affektiv-motivationale und sozio-kulturelle Variablen: Eine empirische Studie*. Kassel, Germany: Kassel University Press.

Holling, H., Böhning, W., Böhning, D., & Formann, A. K. (2016). The covariate-adjusted frequency plot. *Statistical Methods in Medical Research*, *25*, 902–916. https://doi.org/10.1177/0962280212473386

Hulstijn, J. H. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language and Cognition*, *15*, 422-433. https://doi.org/10.1017/S1366728911000678

Hung, L.-F. (2012). A negative binomial regression model for accuracy tests. *Applied Psychological Measurement*, *36*, 88-103. https://doi.org/10.1177/0146621611429548

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, *14*, 47-84. https://doi.org/10.1177/026553229701400104

Larwin, K., & Milton, H. (2012). A demonstration of a systematic item-reduction approach using structural equation modeling. *Practical Assessment, Research, & Evaluation*, *17*(8), 1-19.

Li, E. F. (2013). The impact of unobserved extreme categories on item and person estimates-A simulation study. In Q. Zhang and H. Yang (Eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding* (pp.117-128). Berlin, Germany: Springer.

Linnemann, M., & Wilbert, J. (2010). The C-test: A valid instrument for screening language skills and reading comprehension of children with learning problems?

Field Code Changed

Formatted: English (United States)

In R. Grotjahn (Ed.), *The C-test: Contributions from current research* (pp. 113-124). Frankfurt, Germany: Lang.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174. https://doi.org/10.1007/BF02296272

Mozgalina, A., & Ryshina–pankova, M. (2015). Meeting the challenges of curriculum construction and change: Revision and validity evaluation of a placement test. *The Modern Language Journal*, *99*, 346-370. https://doi.org/10.1111/modl.12217

Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika*, *52*, 165-181. https://doi.org/10.1007/BF02294232

Norris, J. M. (2018). Developing and investigating C-tests in eight languages: Measuring proficiency for research purposes. In J. M. Norris (Ed.), *Developing C-tests for estimating proficiency in foreign language research* (pp. 7-33). Frankfurt, Germany: Lang.

Organization for Economic Cooperation and Development (2017). *PISA 2015 Technical Report*. Paris: OECD. Retrieved from: https://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

Raatz, U., & Klein-Braley, C. (1981). The C-Test—A Modification of the Cloze Procedure. In T. Culhane, C. Klein-Barley, & D. Stevenson (Eds.), *Practice and Problems in Language Testing* (pp.113-148). Clochester, UK: University of Essex Occasional Paper.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Oxford, UK: Nielsen & Lydiche.

Reichert, M., Brunner, M., & Martin, R. (2014). Do test takers with different language backgrounds take the same C-test? The effect of native language on the validity of C-tests. In R. Grotjahn (Ed.), *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends* (pp. 109-135). Frankfurt, Germany: Lang.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika*, *53*, 349-359. https://doi.org/10.1007/BF02294217

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*, 237-247. https://doi.org/10.1111/j.1745-3984.1991.tb00356.x

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston, MA: Allyn & Bacon/Pearson Education.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-70. Https://doi.org/10.1177/109442810031002

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*, 228-243. https://doi.org/10.1037/a0027127

Wockenfuß, Verena & Raatz, Ulrich. (2014). Zur Validität von muttersprachlichen C-Tests: Bedeutung von verbaler Intelligenz und Informationsverarbeitungsgeschwindigkeit unter Berücksichtigung des Lebensalters. In R. Grotjahn (Ed.), *Der C-Test: Aktuelle Tendenzen/The C-test: Current trends* (pp. 191-224). Frankfurt, Germany: Lang.

Field Code Changed

Formatted: English (United States)

Footnotes

[1] A model of weak invariance cannot be formulated for the outlined count data models because discrimination parameters are not included in the models.

[2] Please see the OSF repository (LINK) for a check of the missing data pattern that justifies this exclusion strategy.

Table 1.

Description of the participants.

	Native speakers	Second language learners
N	129	127
Sex		
# of females	105	93
# of males	24	34
Age at testing		
# of missings	0	2
M	25.12	28.55
SD	6.23	6.67
Min	15	16
Max	48	59

Field Code Changed

Formatted: English (United States)

Formatted: English (United States)

Formatted: English (United States)

Field Code Changed

Formatted: English (United States)

Table 2

Correlations of ability estimates and reliability estimates.

		1	2	3	4	5
RSM	1	.961				
PCM	2	.999	.961			
RPCM	3	.976	.967	.943		
CMPCM with global dispersion	4	.973	.965	1.00	.969	
CMPCM with item-specific dispersion	5	.972	.963	.999	.999	.969

Notes. N = 256. All *p*s < .001 for all correlations. Estimates of empirical reliability are depicted on the diagonal (Brown & Croudace, 2015).

Field Code Changed

Formatted: English (United States)

Table 3

Item descriptive statistics.

	M	*SD*	r_{it}	Timelimit (seconds)	timelimit (minutes)
Item 1	15	6.4	.90	100	1.67
Item 2	17	6.9	.92	85	1.42
Item 3	16	7.1	.93	80	1.33
Item 4	15	6.4	.89	115	1.92
Item 5	14	6.2	.90	65	1.08
Item 6	14	6.8	.91	85	1.42

Notes. N = 256; r_{it} = item-scale correlation with part-whole correction; timelimit = time on task in seconds resp. minutes – the logarithm of these times (in minutes) was used as an offset in the RPCM and both CMPCMs.

Field Code Changed

Formatted: English (United States)

Table 4

Item descriptive statistics and model estimation results for RPCM and CMPCMs.

	RPCM	CMPCM with global dispersion	CMPCM with item-specific dispersion
	Fixed effects		
	$\beta\ (SE_\beta)$	$\beta\ (SE_\beta)$	$\beta\ (SE_\beta)$
Item 1	2.092 (0.036)***	2.087 (0.035)***	2.089 (0.035)***
Item 2	2.374 (0.035)***	2.369 (0.035)***	2.371 (0.034)***
Item 3	2.363 (0.036)***	2.358 (0.035)***	2.360 (0.035)***
Item 4	1.981 (0.036)***	1.976 (0.035)***	1.978 (0.035)***
Item 5	2.466 (0.036)***	2.461 (0.035)***	2.463 (0.034)***
Item 6	2.170 (0.036)***	2.166 (0.035)***	2.167 (0.036)***
	Random effect		
σ_θ^2	0.259	0.275	0.270
	Dispersion		
	τ	$\tau\ (SE_\tau)$	$\tau\ (SE_\tau)$
Global	0	-0.658 (0.040)***	-
Item 1	-	-	-0.692 (0.104)***
Item 2	-	-	-0.996 (0.115)***
Item 3	-	-	-0.656 (0.103)***
Item 4	-	-	-0.601 (0.103)***
Item 5	-	-	-0.804 (0.105)***
Item 6	-	-	-0.314 (0.098)***
Model comparison			
$\Delta\chi^2(df)$	-	214.911 (1)***	22.374 (5)***
AIC	8566.674	8353.763	**8341.388**
BIC	8604.033	**8396.458**	8410.768

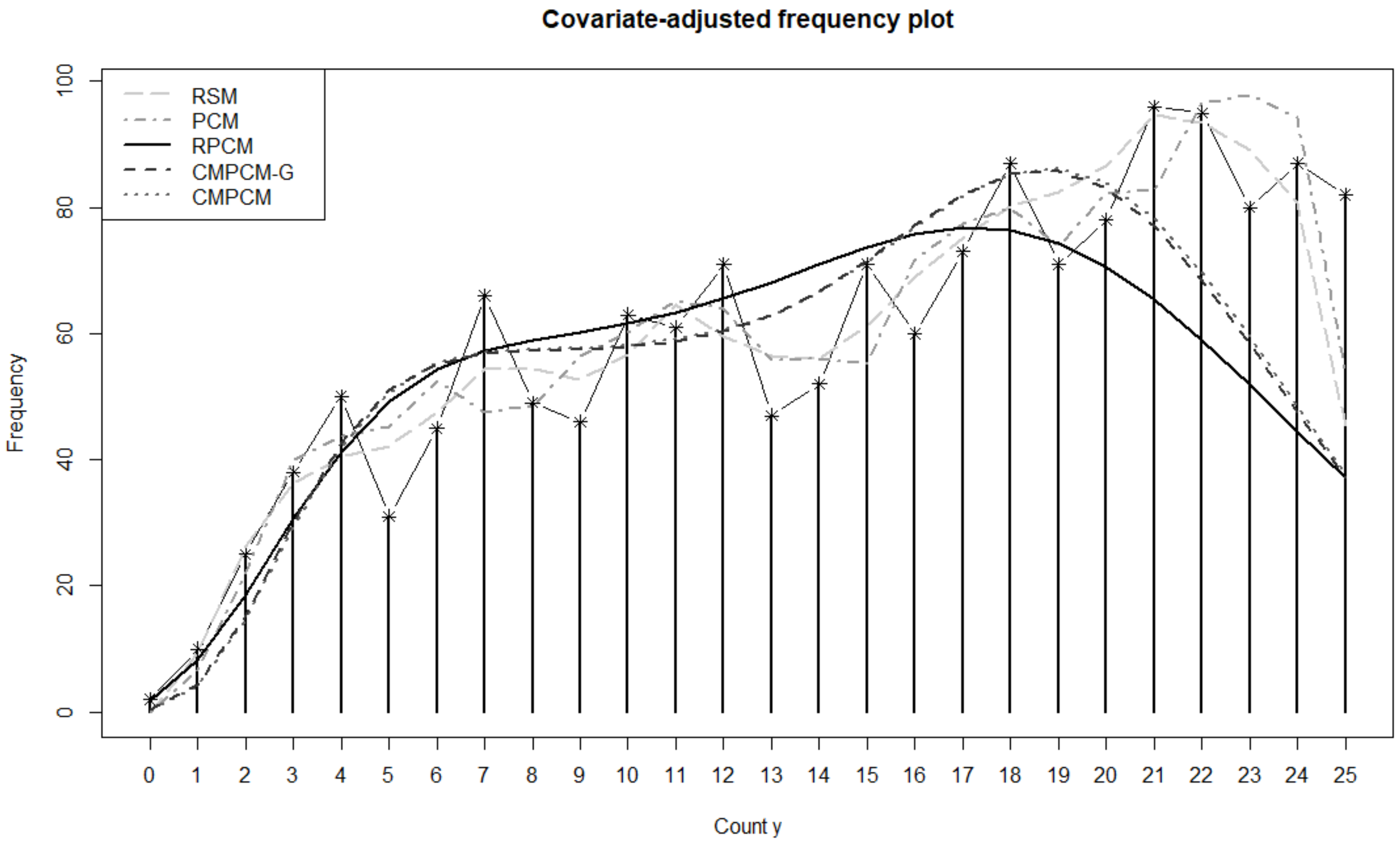
Notes. $N = 256$;The logarithm of the time on task in minutes (see Table 3) was used as an offset in the RPCM and both CMPCMs; $\tau =$ dispersion parameter ($\tau < 0$ indicates underdispersion; $\tau = 0$ indicates equidispersion; and $\tau > 0$ indicates overdispersion). CMPCM with global dispersion is compared with RPCM and CMPCM with item-specific dispersion is compared with CMPCM with global dispersion. $\Delta\chi^2 =$ likelihood ratio statistic; AIC = Akaike’s Information Criterion; BIC = Bayesian Information Criterion. Lower values of information criteria imply better model fit. $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

Field Code Changed

Formatted: English (United States)

Field Code Changed

Formatted: English (United States)



Field Code Changed

Formatted: English (United States)

Figure 1. Covariate-adjusted frequency plot for the RSM, PCM, RPCM, CMPCM with global dispersion (CMPCM-G), and CMPCM with item-specific dispersion (CMPCM). Model-implied counts are depicted by black or gray lines and they should be as close as possible to the observed frequency counts to indicate model fit (Holling et al., 2016).

Field Code Changed

Formatted: English (United States)

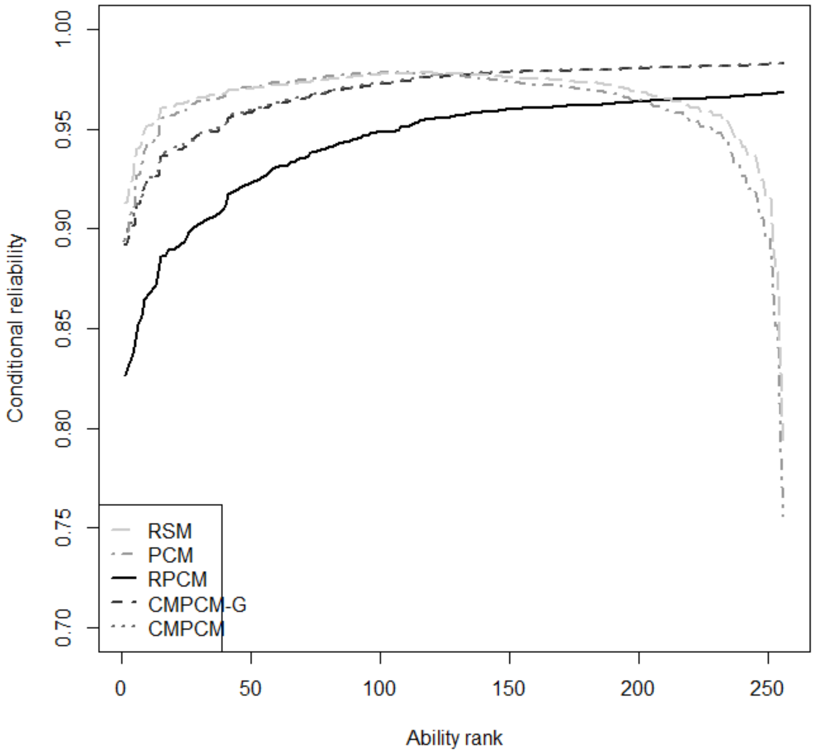
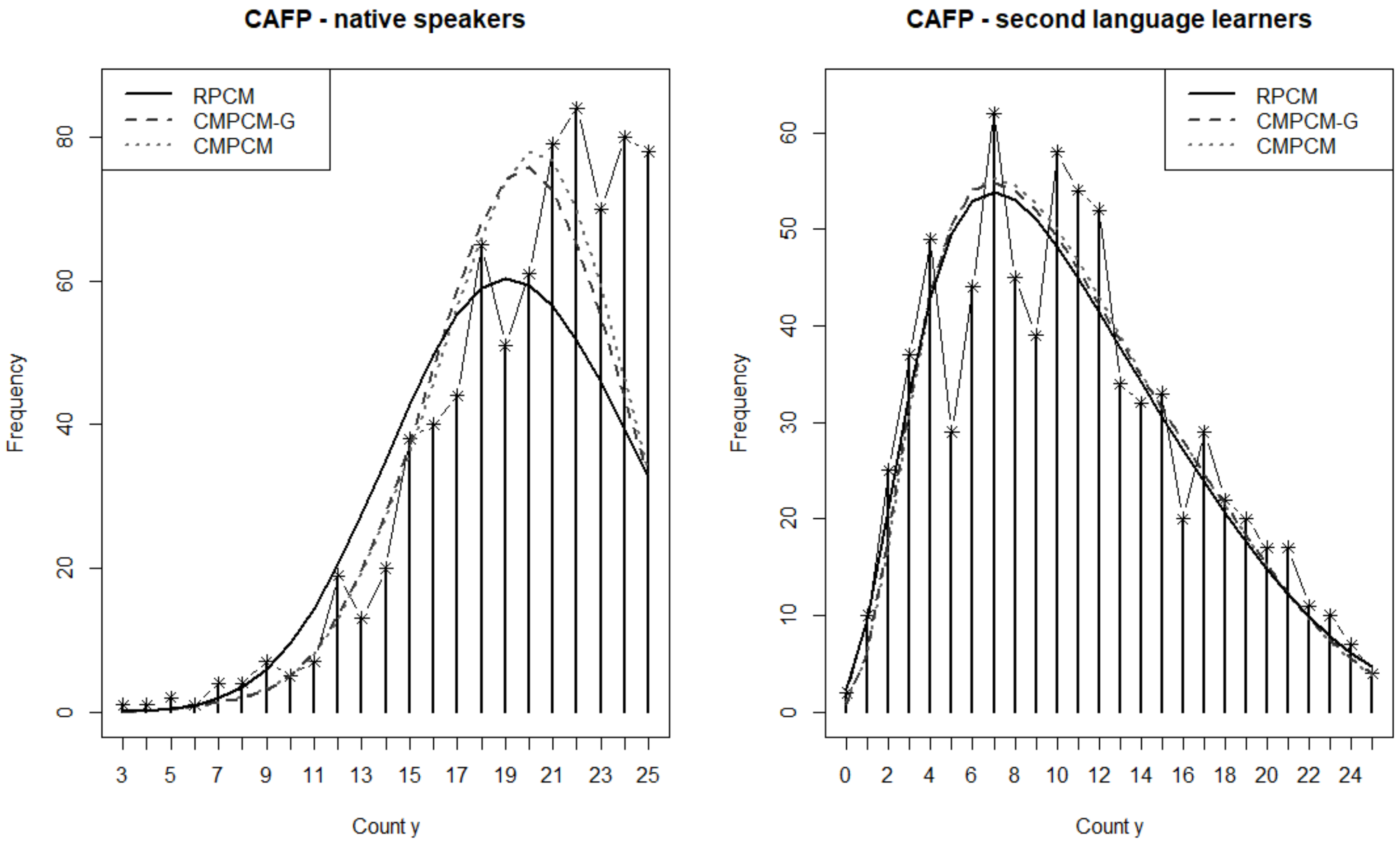


Figure 2. Conditional reliability estimates (y-axis) as a function of ranked ability estimates (x-axis) for RSM, PCM, RPCM, CMPCM with global dispersion (CMPCM-G), and CMPCM with item-specific dispersion (CMPCM).

Field Code Changed

Formatted: English (United States)



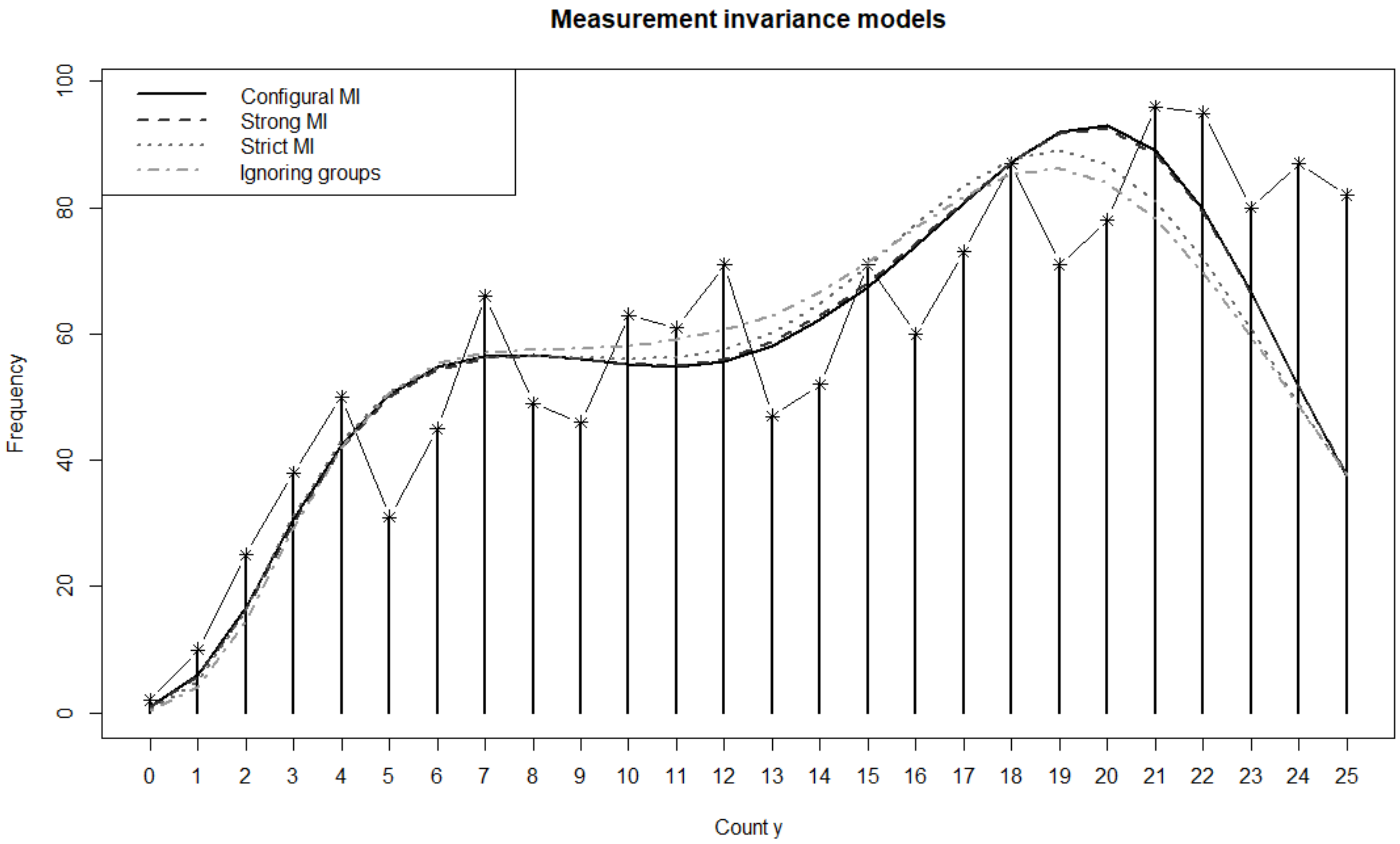
Field Code Changed

Formatted: English (United States)

Figure 3. Covariate-adjusted frequency plot for the RPCM, CMPCM with global dispersion (CMPCM-G), and CMPCM with item-specific dispersion (CMPCM). Model-implied counts are depicted by black or gray lines and they should be as close as possible to the observed frequency counts to indicate model fit (Holling et al., 2016). Left: CAFPs for native speakers. Right: CAFPs for second language learners.

Field Code Changed

Formatted: English (United States)



Field Code Changed

Formatted: English (United States)

Figure 4. Covariate-adjusted frequency plot for the configural invariant model, the strong invariant model, the strict invariance model, and the CMPCM with item-specific dispersion ignoring language level groups. Model-implied counts are depicted by black or gray lines and they should be as close as possible to the observed frequency counts to indicate model fit (Holling et al., 2016).

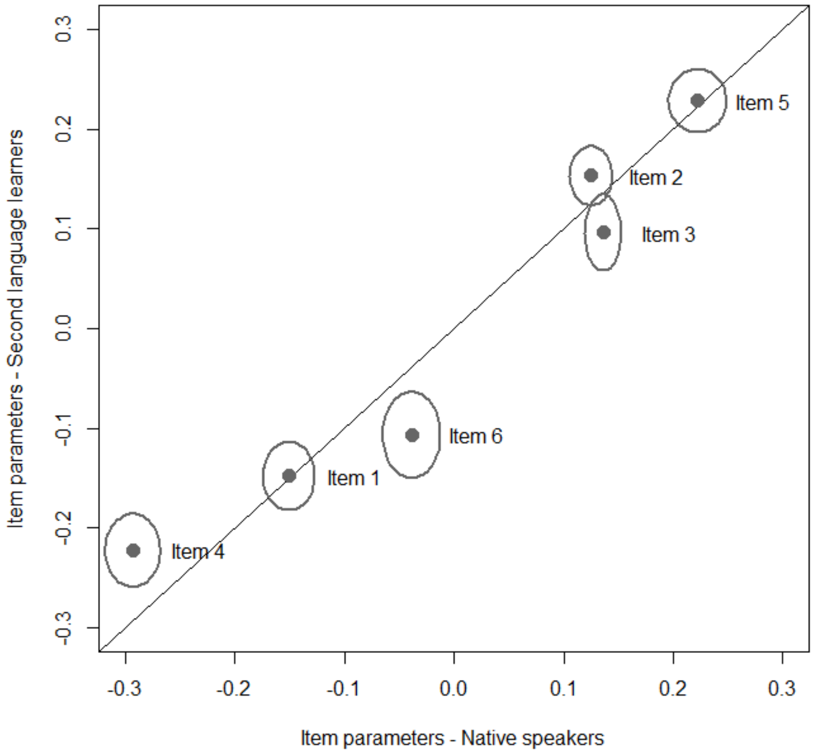


Figure 5. For each item the difference to the average easiness in each of the groups is plotted (centered for both groups). 95% confidence ellipses are plotted around the item parameters: If they cover the reference line, parameters do not significantly differ across groups.