The Effects of the Violation of Local Independence Assumption on the Person Measures under the Rasch Model

Pourya Baghaii Moghadam
Islamic Azad University of Mashhad
&
Reza Pishghadam
Ferdowsi University of Mashhad

Abstract
Local independence of test items is an assumption in all Item Response Theory (IRT) models. That is, the items in a test should not be related to each other. Sharing a common passage, which is prevalent in reading comprehension tests, cloze tests and C-Tests, can be a potential source of local item dependence (LID). It is argued in the literature that LID results in biased parameter estimation and affects the unidimensionality of the test. In this study the effects of the violation of the local independence assumption on the person measures in a C-Test are studied. A C-Test battery comprising four passages, each containing 25 blanks, was analysed twice. Firstly, each gap was treated as an independent item and Rasch's (1960) dichotomous model was employed. In the second analysis, each passage was treated as a super item and Andrich's (1978) rating scale model was used. For each person, two ability measures were estimated, one on the basis of the dichotomous analysis and one on the basis of the polytomous analysis. The differences between the two measures, after being brought onto the same scale, are compared and the implications are discussed.

Key Words: C-Test, local item dependence, Rasch model, rating scale model, separation

Introduction

The Rasch model is a probabilistic model which has its roots in conditional probability and independence of events. In fact, most of the

interesting properties of the Rasch model derive from the probability theory which works when certain assumptions hold. One of the interesting properties of the Rasch model is the separation of person and item parameters. This means that the relative difficulties of items can be estimated independently of persons' abilities. And similarly, the abilities of the persons can be estimated independently of the difficulties of the items. It is algebraically possible to estimate the probability of any response pattern to a set of items by a test-taker. The response patterns which are more on a Guttman scale are more probable and those which are aberrant have lower probabilities. It implies that if an examinee can successfully answer items of that difficulty, s/he would be able to answer the earlier questions.

Consider a two-item test where there are 4 possible patterns of response:
0 0
1 0
0 1
1 1

The first and last patterns are not helpful because when the responses are the same we have no criterion for estimating the relative difficulties of the items. We can determine the relative standing of the items in terms of difficulty only when the responses are different. The probability of the first pattern, i.e., the probability of the first item being right, when the second is wrong is:

$$\text{Prob}\ \{X_1=1, X_2=0\} = e^{-d_1} / e^{-d_1} + e^{-d_2}$$

Where *e* is the logarithmic constant, d_1 is the difficulty of the first item and d_2 is the difficulty of the second item. The probability that the first item is wrong when the second is right is:

$$\text{Prob}\ \{X_1=0, X_2=1\} = e^{-d_2} / e^{-d_1} + e^{-d_2}$$

The important issue about these equations is that they contain no person parameters, which means that the relative difficulties of the items can be estimated without assuming anything about the ability of the

persons who happen to have taken the test. The corollary of the above equations can be derived when two persons respond to one item. That is, the probability that person 1 gets item i right and person 2 gets the same item wrong equals:

Prob $\{X_{1i}=1, X_{2i}=0\} = e^{b_1} / e^{b_1} + e^{b_2}$

Where e is the logarithmic constant, b_1 is person one's ability measure and b_2 is person two's ability measure. As you can see, this equation does not contain any item difficulty parameter which means that the relative abilities of persons can be estimated independently of the difficulties of the items.

The concept above can be put in a slightly different way: having the total score of a person, the response pattern depends only on the difficulties of the items and we do not need any information about persons' ability estimates (extreme scores do not provide any relative information and the probability of these patterns cannot be estimated). This is an extremely important concept in Rasch model which states that the person's total score contains all of the information about person's ability and the total score of an item contains all the information about the difficulty of the item. This concept is referred to as *sufficiency* in the literature, i.e., to estimate person and item parameters, the total score suffices. Of course, this is only true when the responses fit the Rasch model. For each total score, there are several patterns of response depending on the number of items in the test. Each response pattern has a probability of occurring. Those which are closer to the Guttman pattern have higher probabilities. Thus, if the responses fit the Rasch model, we expect most of the patterns to be nearly Guttman-like. However, if there are many persons with response patterns which are not Guttman-like and are estimated to have low probabilities of occurrence, we say that the responses do not fit the Rasch model and the total score is not the sufficient statistic to estimate the person ability (Andrich & Marais, 2006).

Computation of the probability of the occurrence of each response pattern and the nice *separation property* that derives from it comes, however, at a price. This price is the independence assumption that we make for the computation of the probability of the response patterns. Statistically speaking, event X is independent of event Y if the probability of the occurrence of X is not affected by the occurrence of Y. In computing the response patterns, we are applying conditional probability. When we compute the probability of a given response pattern say, {1 0} in our two-item test, we are computing the probability that the first is right under the condition that the second is wrong for a person at a given ability level and for items at given difficulty levels. We can only do this if we assume that reply to item 1 does not influence reply to item 2 and vice versa. This is where the assumption of local independence of items in Rasch model roots. The items that are put to Rasch analysis are required to be independent of each other. That is, a correct or wrong reply to one item should not lead to a correct or wrong reply to another item. This means that there should not be any correlation between two items after the effect of the underlying trait is conditioned out, i.e., the correlation of residuals should be zero. The items should only be correlated through the latent trait that the test is measuring (Lord & Novick, 1968). If there are significant correlations among the items after the contribution of the latent trait is removed, i.e., among the residuals, then the items are locally dependent or there is a subsidiary dimension in the measurement which is not accounted for by the main Rasch dimension (Linacre, 1998; Lee, 2004). In other words, performance on the items depends to some extent on a trait other than the Rasch dimension which is a violation of the assumptions of local independence and unidimensionality.

> If the assumption of local item independence is violated, any statistical analysis based on it would be misleading. Specifically, estimates of the latent variables and item parameters will generally be biased because of model misspecification, which in turn leads to incorrect decisions on subsequent statistical analysis, such as

> testing group differences and correlations between latent variables. In addition, it is not clear what constructs the item responses reflect, and consequently, it is not clear how to combine those responses into a single test score, whether IRT is being used or not (Wang & Wilson., 2005, p.6).

When a set of items are locally dependent, they are usually bundled into polytomous super-items, that is, the set of items which are related to a common stimulus are considered as one polytomous item to partial out the influence of local item dependence (LID) among items within each super-item. Polytomous Rasch models or IRT models such as Andrich's rating scale model (1978) or Master's (1982) partial credit model, etc. are then applied to analyse the test-lets. The drawback of collapsing dichotomies into polytomies, however, is the loss of information. We can only get parameter estimates for the test-lets. Each test-let may contain several dichotomous items about which we obtain no information (Yen, 1993).

It is argued in the literature that if the local independence assumption does not hold, the local dependence itself acts as a dimension. If the effect of LID is substantial, it is difficult to say what dimension the main Rasch dimension is. Even if the effect is small, the derived measures will be contaminated, i.e., the measures partially reflect the LID dimension to the extent that LID exists. In fact, LID is a form of violating the unidimensionality principle. LID also results in the overestimation of reliability and consequently artificially small standard errors of estimates (SEE). This could be a very severe problem in computerized adaptive testing where SEE is the criterion for terminating the test. It can result in premature termination of the test (Zenisky, Hambelton, & Sireci, 2003). The problem of LID is not new and has also been addressed in the classical test theory (CTT).

Dependency among items can inflate reliability and give a fake impression of the precision and quality of the test. One example of the

tests enjoying dependency among items is C-Test. As an alternative to cloze procedure, C-Test is a test consisting of four to six thematically distinct segments of connected discourse, in which the second half of every second word is deleted. C-Test like cloze test seems to violate LID, because if an examinee fails to answer a question, it may impact the other responses too (Klein-Braley, 1985).

As mentioned before, it seems that violating the assumption of LID can affect the results. Therefore, this study attempts to find out whether the violation of the assumption of LID on the person measures in a C-Test can impact decision making in a hypothetical high-stakes assessment.

Methodology

Participants

A community sample of 160 people participated in this study, comprising 61 males and 99 females aged between 19 and 32 (M= 20.1, SD=9.7). All of the participants were university students attending four universities in Iran, Mashad, majoring in English language literature (81), translation (40), and teaching (39) and were at different years of their undergraduate studies.

Instrumentation

Two tests were utilized to conduct this study: a C-Test and a reading test. The C-Test was taken from a test battery developed and validated by Klein-Braley (1985), which included 23 passages with known difficulty levels. For the purposes of this study, the easiest and the most difficult passages with p-values of 0.83 and 0.39 respectively, were chosen as the first and last super-item in the battery. Two other passages with p-values of 0.76 and 0.65 were selected as the second and third super-items. Each passage had 25 blanks. The Cronbach's Alpha reliability of the C-Test was 0.88.

Alongside the C-Test battery, a reading comprehension test was also administered. It was one of the samples of Cambridge CAE practice

tests published by Cambridge University Press (1991). The test consisted of two parts. In the first part, there was a long passage followed by eight multiple choice questions. In the second part, there was another passage with six missing sentences. The missing sentences were on another page with an extra sentence. The students were required to find out where in the text the missing sentences fit.

The Cronbach's Alpha reliability of the reading test was 0.81. The correlation between the two sections of the reading test i.e., the multiple choice (MC) test and the sentence insertion (SI) was 0.65 which is significant ($p<0.01$, $N=160$), and the entire reading test and the C-Test significantly correlate at 0.53 ($p<0.01$, $N=160$).

To substantiate the validity of the reading test, it was Rasch analysed. Table 1 shows the relevant Rasch statistics. Except for SI5 which misfits, almost all items show good fit. Some of the Zstd statistics are out of the acceptable –2 to +2 range, but since all the infit mean squares are within the right range of 0.75-1.30, all the items except SI5 were considered to be measuring one unidimensional latent variable. It should be noticed that all the out of range infit Zstd except one excess –2 which is an indication of overfit. The overfit of some of the items was expected due to the local dependence of the items. The point-measure (PTMEA) correlation is the Rasch analogue of point-biserial correlation in CTT. As Table 1 shows, they are all noticeably positive.

Table 1
Measure order and fit statistics for reading items

Entry raw	Model	Infit	Outfit	PTMEA	Number Score Measure S.E.	MNSQ ZSTD	MNSQ ZSTD	Cor.	Item
13	15	2.64	.29	1.15	.8	2.44	2.2	.29	SI5
10	36	1.35	.22	1.01	.1	1.05	.3	.49	SI2
9	41	1.12	.21	1.10	1.0	1.08	.4	.46	SI1
6	45	.94	.21	.99	- .1	.99	.0	.52	RC6
2	56	.50	.20	1.21	2.2	1.28	1.6	.43	RC2
5	63	.23	.19	1.03	.4	1.13	.9	.52	RC5
14	67	.08	.19	.81	-2.4	.70	- 2.4	.64	SI6
11	86	-.62	.19	.82	-2.2	.72	- 2.1	.64	SI3
1	88	-.70	.19	1.08	1.0	1.10	.7	.51	RC1
7	88	-.70	.19	.98	- .2	.99	.0	.56	RC7
3	92	-.85	.20	.92	- .9	.86	- .9	.59	RC3
12	98	-1.08	.20	.78	-2.6	.64	- 2.3	.66	SI4
4	102	-1.24	.20	1.10	1.1	1.02	.2	.50	RC4
8	112	-1.67	.21	.96	- .3	1.16	.7	.53	RC8
mean	70.6	.00	.21	.99	- .2	1.08	.0		
S.D.	27.8	.1.16	.03	.13	1.4	.42	1.3		

Data Collection and Analysis

To collect the data, the participants were firstly asked if they would volunteer to take part in a study on "language profcicency and C-Test". Then, getting permission from the professors of the universities, in September (2007), at the start of the academic year, participants were asked to take the tests, i.e. the C-Test and reading test.

Scoring the tests, the researchers analysed the data twice, once using Rasch's (1960) dichotomous model treating each gap as an independent dichotomous item, and once treating each passage as a ploytomous item or test-let using Andrich's (1978) rating scale model. So, for each person, two measures were obtained, one based on the dichotomous analysis and one based on the polytomous analysis. Afterwards, the subjects were divided into 3 ability groups on the basis

of their C-Test scores, and the correlations between the C-Test and the reading test were computed for each ability group separately.

Results and Discussion

The table below shows the descriptive statistics for the sample from the two analyses after being brought onto the same scale.

Table 2
Descriptive statistics for person measures from the two analyses

Analysis	N	Minimum	Maximum	Mean	S. D	Variance
Dichotomous	160	-3.4	1.9	-.346	1.2417	1.542
Polytomous	160	-2.99	1.51	-.331	1.06111	1.126
Valid N	160					

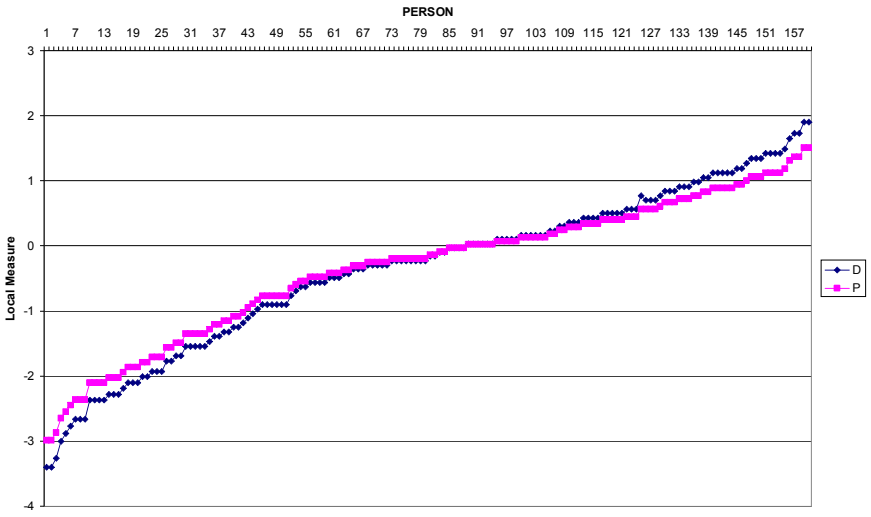
As Table 2 indicates, there is not much difference in the measures obtained from the two analyses. The means are almost the same and the measures correlate at 1. The dichotomous analysis; however, gives a wider spread of measures. The contrasts, i.e., the differences between measures from the two analyses range between 0.00-0.41 logit and are not significant. These differences are negligible and very close to the standard error of the estimates. That is, the measures from the dichotomous analysis lie within one standard error of the estimate of the polytomous analysis.

The measures from the two analyses vary in a peculiar way. At the two ends of the ability scale the measures are different and at the middle they are almost the same. That is, the type of the analysis affects persons who have obtained either low measures or high measures, and the middle ability test-takers are not affected by the type of the analysis.

Figures 1 and 2 show the location of each individual on the (vertical) ability scale based on the two analyses. The persons are on the horizontal axis and have been ordered by their ability estimates, i.e., person 1 is the least able student and person 160 is the most able one.

Interestingly, as the left bottom corner of the plot shows, the polytomous analysis favors low ability students as its corresponding line is above the dichotomous analysis line. For the mid-ability students, they almost overlap as the two lines become one and for the high ability students, as the top right corner of the plot indicates, the dichotomous analysis gives higher measures than the polytomous analysis.

Figure 1
Plot of person measures from the two analyses

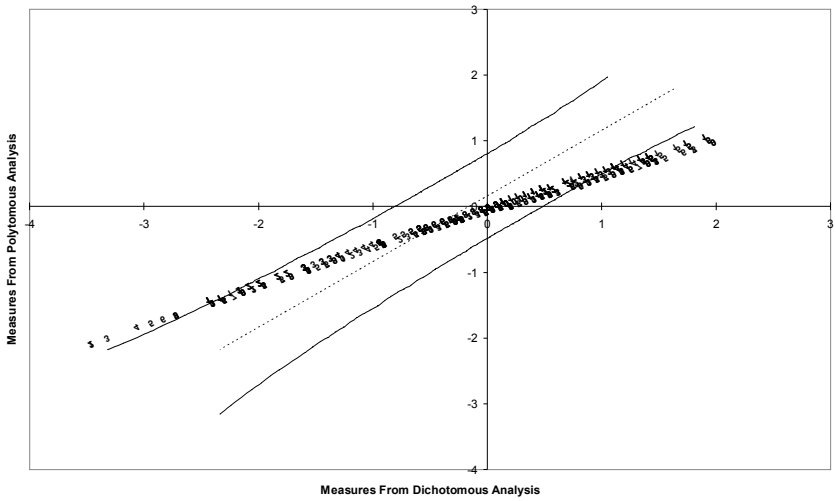


The Cronbach's Alpha reliability of the test when all the 100 gaps are considered as items is 0.95. When the 4 passages are entered into the reliability analysis as 4 polytomous items the reliability drops to 0.88. This drop in reliability is expected because the number of items is reduced from 100 to only 4. Another reason for the drop in the reliability is the lack of local dependence in the polytomous analysis since local dependence can inflate alpha. Here, when the gaps are considered as items, the reliability augments to 0.95 from 0.88, i.e., 0.07 of increase. However, it is not clear to what extent this increase is due to

local dependence and to what extent it is due to increase in the number of items. According to the Spearman-Brown Prophecy formula, in order to boost the reliability of the test from 0.88 to 0.95, we need to lengthen the test 2.59 times. In this study, the test was lengthened 25 times, much more than was needed. Therefore, there is no way to disentangle the effect of increase in test length from the effect of local dependence (as a result of dichotomous analysis) on the reliability boost in the case of these data. The question that may arise here is that: why doesn't the reliability of the test increase more than 0.95 since we lengthened the test around 10 times more than required according to the Spearman-Brown Prophecy formula? The answer is that lengthening the test improves the reliability to a certain point beyond which the length has an insignificant effect on reliability.

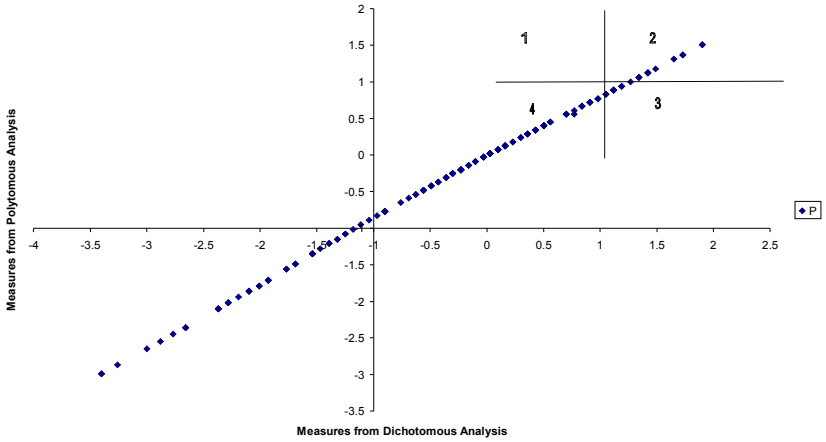
Although the existence of local dependence in the C-Test is obvious, the aforementioned analysis reveals that the measures from the two analyses are very close and correlate at 1. Therefore, if norm-referenced decisions are made, it does not really matter which analysis is used as the basis of decision making. The largest difference in the case of the present data is 0.40 logits. The cross plot of the person measures and 95% quality control lines show that several of the extreme scorers fall outside the quality control lines. This means that the measures obtained from the two analyses for these students are not within modeled measurement error (Wright & Stone, 1979).

Figure 2
Scatter plot of person measures from the two analyses with 95% control lines



As far as criterion-referenced decision-making is concerned, we do make adverse decisions depending on which analysis we use. In the following plot, a hypothetical cut-score at +1 logit is imposed. For persons who fall in areas 2 and 4, the same decisions will be made. Test-takers who fall in areas 1 and 3 would have opposite decisions, depending on the analysis. Here, no one falls in area 1 but four test-takers fall in area 3. That is, if we base our decision-making on the dichotomous analysis, these four people pass and if we decide on the basis of polytomous analysis, these four test-takers fail. Therefore, by using dichotomous analysis, four people are mistakenly given some privileges that they do not deserve.

Figure 3
Plot of person measures from the two analyses with the hypothetical cut-off score



To corroborate the results more and to see whether the contextual clues vary with the proficiency levels of the testees, along with the C-Test battery, the subjects took a reading test which was intended to be a test of text-level skills. As the following table shows, the correlations increase with gain in proficiency. That is, as the subjects become more proficient, they can make better use of the contextual clues.

Table 3
Correlations between the C-Test and the reading test for the 3 ability groups

	Low-Ability Group	Mid-Ability Group	High-Ability Group
Raw Score Range on the C-Test	9-30	31-50	51-77
Correlation	-0.09	0.19	0.39
Group Size	37	51	72

Conclusions

The results of the study showed that the low-ability students have lower measures on the dichotomous analysis, apparently because they cannot

take advantage of the context clues because of their low proficiency and even what we call context clues here work against them because they have lower measures on the dichotomous analysis. It seems that local dependency negatively covaries with the Rasch dimension for these students. For the middle-ability students in this sample, the two analyses give almost identical measures, meaning that the subjects still are not at a level to benefit from the context though context does not work at their disadvantage. And finally, the high-ability students who can use and benefit from the context clues obtain better measures in the dichotomous analysis. The prima facie evidence shows that for answering C-Test items a *threshold effect* is at work. That is, in order for the test-takers to be able to take advantage of the context, they need to be at a certain ability level, below which context does not help.

The results obtained as to the purpose of the study can hopefully be interpreted as having some implications for testers. As the results of this study demonstrate, the scores obtained through C-Tests should be evaluated gingerly: how C-test marks are calculated makes no difference for most test takers, but for the highest and lowest scores, the method matters, with a dichotomous analysis favoring high-proficiency and polytomous analysis low-proficiency students. Moreover, when one applies the criterion-reference testing, local dependence affects our interpretation and decision making in a way that some testees are given privileges they don`t deserve; it means some students may pass or fail the course when they shouldn`t. Therefore, testers are advised to avoid any jejune generalization regarding their own findings.

As is clear from any scientific research, nothing can be self evident unless verified by observation or experimentation. To do any type of observation or experiment, one may face with some limitations and problems. This study could have come to somewhat more different results than it did, if it were not confronted with the following limitations. Since in this study the sample was not much large, a replication of this study with a sample of wider ability spread in both directions could be very informative. It would be interesting to see

whether the use of the contextual clues monotonically increases with ability or it stops at some ability location or even drops. That is, would the two lines diverge more sharply at the two corners of the plot had we tested subjects with more extreme scores or would they overlap at some stage or their locations change? Finally, other results would be obtained if this study were triangulated with other types of dependency-based instruments like cloze or reading comprehension tests.

Received 5 January, 2008
Accepted 23 March, 2008

Acknowledgements
We gratefully acknowledge the project reported here was supported by a grant-in-aid of research from the Iran `s National Institute of Elites to the second author without which this research would not have been possible.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.

Andrich, D. & Marais, I. (2006). *Instrument design with Rasch IRT & data analysis 1*. *Unit Materials, Semester 2, 2006*. Perth, Australia: Murdoch University Press.

Aspinall, P. & Hashemi, L. (1991). *CAE Practice Tests 1*. Cambridge: Cambridge University Press.

Klein-Braley, C. (1985). A cloze-up on the C-Test: A study on the construct validation of authentic tests. *Language Testing,* 2, 76-104.

Lee, Y. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing*, 21, 74-100.

Linacre, J. M. (1998). Structure in Rasch residuals: why principal component analysis? *Rasch measurement transactions*, 21(2), 636. Retrieved from: *http://www.rasch.org/rmt/rmt122m.htm*

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47*,* 149-174.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960. (Expanded edition, Chicago: The university of Chicago Press, 1980).

Wang, W. & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, 29, 296-318.

Wright, B. D. & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement,* 30, 187–213.

Zenisky, A. L., Hambelton, R. K. & Sireci, S. G. (2003). Effects of local item dependence on the validity of IRT item, test and ability statistics. *Association of American Medical Colleges (AAMC).* Retrieved from: *http://www.aamc.org/students/mcat/research/monograph5.pdf*