Nested Factor Analytic Model Comparison as a Means to Detect Aberrant Response Patterns

John M. Clark III

Pearson

Author Note

John M. Clark III, Psychometric and Research Services, Pearson.

Correspondence concerning this manuscript should be addressed to Mike Clark, Pearson, 2488 E. 81st St., Suite 4700, Tulsa, OK. E-mail: mike.clark@pearson.com

Abstract

Although specific methodologies vary across the numerous examples of person-fit statistics, in general, they work by measuring the extent to which an individual's observed response vector is consistent with expectation as defined in the context of some statistical model. Response patterns that are inconsistent with expectation are said to be aberrant, while responses that are consistent with expectation are understood to be non-aberrant. This paper discusses the results of a small simulation study designed to investigate a proposed method for identifying person-level misfit not by evaluating the extent to which the individual's response pattern conforms to expectation as defined by the desired statistical model, but rather by comparing person-level changes in fit across nested factor analytic models. The rationale behind this proposed procedure is the proposal that cheating behaviors that are intended to be detected by person-fit statistics may introduce multidimensionality into the individual's response pattern, which should be estimable using a two-factor model, and this two-factor model should show significant improvement in person-level fit for individuals who have engaged in cheating behavior. Implications, limitations, and future directions are discussed.

Nested Factor Analytic Model Comparison as a Means to Detect Aberrant Response Patterns

At their most fundamental level, person-fit statistics provide a means to evaluate the extent to which individual response vectors align with expectation as defined by some model. Much like model-fit indices, many residual-based person-fit statistics work by comparing characteristics of observed data to their expected values, with good fit indicated by close alignment between observed and expected values and poor fit indicated by observed values deviating from expectation. Person-fit statistics differ from model fit statistics in that model fit statistics render a single, overall estimate of fit that summarizes the extent to which the entire data set is consistent with expectation, whereas person-fit statistics quantify fit at the individual level, with each test-taker receiving a unique estimate of the extent to which his or her response vector aligns with expectation.

Item response theory provides a useful measurement framework for estimating differences between expected and observed level of performance by virtue of the item response function, which provides conditional probabilities of success on each item given characteristics of the item (i.e., difficulty, discrimination, and lower asymptote, depending on the chosen model) and the individual's ability level. By relating the person's ability level to the characteristics of each item, it is simple to obtain predicted values for the individual for each item, which can then be compared with the individual's observed response vector.

The l_0 person-fit statistic provides a straightforward example of a method that compares observed responses with expected values, as defined by the parameters estimated in an IRT model. In the three parameter logistic (3PL) IRT model, l_0 is estimated as

$$l_0 = \sum_{j=1}^{n} \left\{ X_{ij} \ln P_j(\theta_i) + (1 - X_{ij}) \ln[1 - P_j(\theta_i)] \right\}, \quad (1)$$

where X_{ij} is the observed response of person i to item j, numbered $j = 1, 2, \ldots, n$; θ_i is the

examinee's ability level; and $P_j(\theta_i)$ is the examinee's probability of success on item X_{ij}, given the characteristics of the item and the examinee's ability level. When parameter estimates are inserted into equation (1), l_0 is equal to the value of the test-taker's log-likelihood function evaluated at $\hat{\theta}$ (e.g., Meijer & Sijtsma, 2001). The log-likelihood function—which is used to estimate test-takers' ability parameters—ranges between 0 and negative infinity. One outcome of aberrant response patterns is a flatter log-likelihood function that is less close to the zero when compared to log-likelihood functions computed from non-aberrant response patterns. Recognizing that aberrant response patterns have flatter log-likelihood functions, l_0 was proposed as a method to identify poor person-fit in individuals' response vectors; however, this statistic has a noteworthy limitation. The l_0 person-fit statistic is asymptotically (in terms of test length) referred to the normal distribution. Because tests of infinite length are problematic in real-world applications, the usefulness of l_0 for detecting aberrant response patterns in actual test data is severely limited. In an attempt to create a form of the l_0 statistic with a known sampling distribution, the l_z person-fit statistic was developed (Drasgow, Levine, & Williams, 1985). The l_z statistic was originally thought to follow a standard normal sampling distribution, so person-fit could be evaluated using l_z and z-score flagging criteria. Later studies investigating the characteristics of the l_z statistic have found that it is not normally distributed as originally thought when $\hat{\theta}$ is substituted for θ in computing the statistic (e.g., Nering, 1995; Schmitt, Chan, Sacco, McFarland, & Jennings, 1999; Snijders, 2001).

In a 2007 paper, Ferrando discusses a statistic, referred to as the *lco* scalability index, as a means to evaluate person-fit in the context of factor analysis. Using Ferrando's notation, the observed response of person i to item j (X_{ij}) is equal to

$$X_{ij} = \mu_j + \lambda_j \theta_i + \varepsilon_{ij}, \tag{2}$$

and the expected value of X_{ij} is equal to $\mu_j + \lambda_j \theta_i$ with constant variance $\sigma^2_{\varepsilon j}$ (Ferrando, 2007). In the unidimensional factor model, Ferrando (2007) defines *lco* as

$$lco_i = \sum_{j=1}^{n} \left[\frac{X_{ij} - \mu_j - \lambda_j \theta_i}{\sigma_{\varepsilon j}} \right]^2, \tag{3}$$

where μ, λ, and σ^2_{ε} are item parameters from the factor analytic model and θ is a factor score. This index is a sum of squared, standardized residuals. It then follows that *lco* should be distributed χ^2, with df = n. Ferrando (2007) states that when estimates are used in place of population parameters in the computation of *lco*, it is distributed χ^2 with $n - 1$ degrees of freedom. As discussed by Ferrando (2007), the *lco* scalability index is analogous to the l_0 person-fit statistic, with the useful addition of *lco* following a known sampling distribution, which makes it more suitable than l_0 for measuring person-fit using traditional hypothesis tests.

The *lco* scalability index, as described, is appropriate for one dimensional factor analytic models. In a subsequent paper, Ferrando (2009) described a generalization of the *lco* scalability index to accommodate models incorporating k factors: the *M-lco* scalability index, which is computed as

$$M - lco_i = \sum_{j=1}^{n} \left[\frac{X_{ij} - \mu_j - \lambda_{j1}\theta_{i1} \ldots - \lambda_{jk}\theta_{ik}}{\sigma_{\varepsilon j}} \right]^2, \tag{4}$$

and is distributed χ^2 with $n - k$ degrees of freedom when estimates are used in its computation.

Person-fit statistics, including the limited sample discussed in the present paper and the much larger collection of such statistics that exist in the literature, serve a common purpose: to measure the extent to which an observed response vector conforms to expectation. Response patterns that conform to expectation are said to show good fit and response patterns that do not

are said to show poor fit, but reframing the question of "does the model fit the individual's data?" to become "does a more complex model provide superior fit for the individual's data?" may lead to the uncovering of valuable information when attempting to identify possible incidences of cheating.

In the context of testing, unidimensional measurement models typically are assumed, with each item's total variance being partitioned into a reliable component associated with a single latent construct that is common to all items and a separate random error component. Due to the local independence assumption, items are expected to function independently of one another after accounting for the common component. However, in certain situations related to cheating, there may be reason to expect that other sources of covariance between items—unaccounted for by the default model's single latent construct—may emerge and violate the local independence assumption. For example, should a portion of the items on a test become exposed, and some test-takers take the test with prior knowledge of these items, these test-takers may perform better than expected on these exposed items depending on the characteristics of the individuals and the exposed items.

One approach to identify test-takers who encountered these exposed items is to use a person-fit statistic, which will compare the test-takers' observed performance on the items to their expected levels of performance. An alternative approach would be to fit two competing models, the first a one-factor model (which is consistent in terms of dimensionality with the assumed measurement model for the test) and the second a two-factor model, and test for changes in person-fit across the nested models. The rationale behind this nested model comparison is grounded in the expectation that prior exposure to items, and the improved performance on these exposed items associated with this exposure, will violate local

independence and create additional observed covariance between the exposed items that is inadequately accounted for by the assumed unidimensional model. When a second factor is added to the model, the additional covariance created by the exposure will be modeled by the exposed items' loadings onto the secondary factor. For test-takers who had no exposure to any items prior to testing, a trivial level of improvement in person-fit for these test-takers should be observed in comparing changes in person-fit when a second factor is added to the model. For test-takers who did have contact with exposed items prior to taking the test, there should be a statistically-significant improvement in person-fit when the second factor is included in the model.

The previously-discussed person-fit statistics for factor-analytic models, *lco* and *M-lco*, provide a means to perform statistical tests on changes in person-fit across nested models, although they have not hitherto been used for such purposes. Although the *M-lco* statistic was developed as a means to assess person-fit for factor analytic models with two or more factors, the steps as described by Ferrando (2009) are quite similar to how person-fit is assessed in traditional unidimensional testing scenarios: the fit of each individual's observed response pattern is compared to expectation as defined by the factor analytic model, with poor fit indicated by large values of *M-lco* (with flagging criteria varying depending on the number of degrees of freedom). If both *lco* and *M-lco* are each distributed χ^2 with $n - k$ degrees of freedom when estimates are substituted into equations (3) and (4), then *M-lco* computed from a two-factor exploratory model should be distributed χ^2 with $n - 2$ degrees of freedom, *lco* computed from a one-factor exploratory model should be distributed χ^2 with $n - 1$ degrees of freedom, and their difference should be distributed χ^2 with 1 degree of freedom. Statistically-significant values of this proposed *lco* difference method indicate significant improvement in person-fit with the

addition of a second factor to the model—a possible indicator of cheating.

Method

Simulation Methodology

A small simulation study was conducted to investigate the efficacy of the proposed *lco* difference method for detecting incidences of cheating. The test length used in this study was 15 items, with each simulated item having five score categories. Item responses were simulated from the graded response model using WinGen (Han, 2007), with true *b* parameters distributed with a mean of 0 and standard deviation of 1 and true *a* parameters uniformly distributed between 0.5 and 2.0 with the *D* scaling constant included, and 5,000 examinees were simulated from a true θ distribution with a mean of 0 and standard deviation of 1. Item responses were simulated for 500 replications for each condition, given the true item and person parameters.

To simulate prior exposure for a subset of items and examinees, items were ordered by their true difficulty (based on the values of their expected score functions evaluated at 67% of the maximum possible score of 4) and examinees were sorted by their true θ levels. Following initial simulation of item responses as previously described, rendered item responses were manipulated in accordance with specific characteristics of each experimental condition. The research design for the present study is a 2 × 2 × 3 factorial with the following factors: number of exposed items (3, 6), the number of simulated cheaters in the total data set (51, or approximately 1% of the total number of test-takers; 501, or approximately 10% of the total number of test-takers), and the difficulty of exposed items (only easy items exposed; only difficult items exposed; a combination of easy, moderate, and difficult items exposed). Simulated test-takers chosen to serve as "cheaters" in the data set were selected in equal numbers from three points on the ability distribution. One-third of cheaters were the simulated test-takers with the lowest true ability

levels, one-third were just below the 50th percentile of the ability distribution, and the remaining one-third being selected near the 25th percentile of the distribution. Exposed items were chosen based on their ordered difficulty levels. In the "easy" conditions, either the three or six (depending on the number of exposed items in a given condition) least difficult items were selected to be exposed. In the "hard" conditions, either the three or six most difficult items were selected to be exposed. In the "spread difficulty" condition, the least difficult items, the most difficult items, and the items at the midpoint of the difficulty distribution were chosen to serve as exposed items. In conditions with three exposed items, a single easy, moderate, and difficult item were selected in the "spread" condition. In conditions with six exposed items, two items each were selected from these regions.

After identifying cheaters and exposed items, cheating was simulated by comparing test-taker status (cheater versus non-cheater) and item status (exposed versus not exposed). In instances where simulated cheaters encountered simulated exposed items, a random number was drawn, which resulted in the originally-simulated item response having a 0.90 probability of being recoded to the maximum possible value of 4. If the random draw did not result in the item response being recoded to 4, the originally-simulated value was left unchanged. This process was repeated independently for each instance of a simulated cheater encountering a simulated exposed item.

In addition to the 12 previously-described conditions, an additional condition was included, which included item responses simulated from the same true population parameters that were used in the 12 experimental conditions, but with no additional manipulations performed on item responses from this condition. This condition was included as a means to investigate the distribution of the *lco* difference values and to assess Type I error rates.

Analysis

Following data simulation, exploratory factor analytic models were fit to the manipulated data sets. For each data set, both a 1 and 2-factor exploratory factor analysis model was fit to the data using Mplus with robust maximum likelihood estimation for both models and quartimin oblique rotation for loadings from the two-factor model. For each simulated examinee, *lco*, *M-lco*, and their difference were estimated. For the purposes of recording hit rates and Type I error rates, an alpha level of $\alpha = 0.05$ was selected for the hypothesis tests used in this study. Simulated test-takers with *lco* values greater than 23.68 (i.e., based on χ^2 with df = 14) were flagged, and simulated test-takers with *lco* difference values greater than 3.84 (i.e., based on χ^2 with df = 1) were flagged as well. Flagging was performed on the *lco* statistic as a means to compare and contrast the novel method proposed in this study with a more traditional approach to estimating person-fit.

Results

All one- and two-factor models converged. All models showed good fit, as evidenced by investigation of RMSEA, CFI, and TLI, and no improper solutions (e.g., Heywood cases) were observed. Percentages of simulated cheaters correctly flagged as such by the person-fit statistic (i.e., hit rates) and percentages of test-takers not simulated to be cheaters that were incorrectly flagged by the person-fit statistic (i.e., Type I errors) are provided in Table 1. As shown in this table, the *lco* difference method outperformed the more traditional *lco* method in terms of hit rates in 9 out of 12 experimental conditions. In terms of Type I error rates, both methods flagged slightly more than the expected 5% of simulated test-takers in the control condition in which no cheating behavior was simulated. Type I error rates varied across experimental conditions for both methods.

Discussion

Performance Comparison

Generally speaking, the *lco* difference method outperformed the more traditional *lco* method in terms of correctly identifying simulated cheaters. Figure 1 illustrates some interesting trends observed in the hit rates for these two methods. As shown in this figure, the three graphs in the left column plot hit rates from conditions in which approximately 1% of the sample of 5,000 test-takers within a given data set were simulated to be cheaters, and the graphs in the right column plot hit rates from conditions in which approximately 10% of the sample were simulated to be cheaters. When comparing performance of the *lco* statistic across the 1% and 10% conditions, hit rates decline somewhat uniformly, suggesting that *lco* is less successful in identifying instances of cheating when such behavior is more widespread. This observation that larger proportions of cheaters result in a reduction of power for a person-fit statistic such as *lco* is not altogether surprising, considering that *lco* is a residual-based person-fit statistic.

Because person-fit statistics like *lco* measure the difference between observed and expected performance on an item, but the difficulty of the items—which influences the expected scores—is computed from observed responses, larger proportions of cheaters influence the estimated difficulty of items, so when larger proportions of cheaters are present, their influence lowers the estimated difficulty for exposed items, which results in smaller residuals when comparing observed versus expected performance on items, and therefore reduces the power of residual-based person fit statistics like *lco*. Contrasting that finding with results from the *lco* difference method, some reduction in power was observed in conditions where either only easy items or difficult items were simulated to be exposed, but the *lco* difference method appears to be more robust, with a smaller reduction in power than what was observed for the *lco* statistic.

Increasing the proportion of cheaters actually improved performance of the *lco* difference method when exposed items' had a spread-out range of difficulty, because the additional cheaters helped the model to better estimate a second factor, which will be discussed in further detail shortly.

For both methods, power increased in all conditions when the number of exposed items was increased from 3 to 6. This finding was not surprising, because more exposed items should be beneficial from a detection standpoint for both methods. For the *lco* statistic, a larger proportion of exposed items allows for more opportunities to observe large differences between observed and expected performance—although, similar to what was observed with increasing in the proportion of simulated cheaters—a "tipping point" will inevitably be reached when increasing the proportion of exposed items on a test, and the influence of the large number of exposed items on cheaters' estimated ability levels will make them more difficult to detect. For the *lco* difference method, having 6 exposed items as opposed to 3 makes it easier to estimate a second factor for the exposed items, so the observed increase in power is not unexpected.

In only one condition was the performance of the *lco* difference method extremely poor, and that condition was the one in which approximately 1% of test-takers were simulated to be cheaters, 3 out of 15 items were simulated to be exposed, and the difficulty of exposed items was spread out. Only 12.5% of simulated cheaters in this condition were correctly identified by the *lco* difference method, compared with a 52.0% hit rate from the *lco* statistic, which is not itself incredibly impressive, but an improvement nonetheless. Upon closer examination of rotated factor loadings from the two-factor model, it became apparent that in this particular condition—with very few simulated cheaters, few simulated exposed items, and exposed items having very different difficulty levels from one another—that a second factor failed to emerge, which is why

so little power was observed in this condition. Power improved dramatically with either the addition of more exposed items or more cheaters because a much more prominent second factor emerged in both cases.

Limitations

Practical considerations. The *lco* difference method described in this paper represents one manifestation of a promising new direction in assessing person-fit, but its potential for use in practical testing applications is currently limited. With future changes coming by way of next generation assessments, there may come a day in which tests comprised entirely of polytomously-scored items with sufficient score categories to be included in factor analytic models become a reality, but for the time being, multiple-choice items dominate the testing landscape. The method described in the present paper is not well-suited to assess person-fit for dichotomously-scored items. Most common applications of factor analysis—those that use ML estimation in particular—assume that observed variables are normally distributed. Dichotomous variables violate that assumption. Of course, there are well-documented ways to get around the dichotomous indicator variable problem—most often by estimating a tetrachoric correlation matrix from the dichotomous response data and using either a weighted least squares or a modified weighted least squares estimation method for the model (e.g., Brown, 2006). The issue that remains is that the *lco* and *M-lco* statistics make a rather important assumption that each item's residual variance from the factor analysis model accurately estimates the error associated with observed scores. Because both *lco* and *M-lco* use the square root of the item's FA residual variance estimate to compute standardized residuals, which are then squared and summed to compute the statistic, if the items' residual variance estimates from the model do not adequately represent error variance, then the residuals are not properly standardized, and therefore the

resulting distribution of the *lco* or *M-lco* statistic is not χ^2 with df = $n - k$. This issue was observed in preliminary attempts to apply this method to test comprised of dichotomous data and requires further research in the future.

Simulation methodology. Simulations are useful for research such as this because they provide both an opportunity to control desired characteristics of the data and because they allow true characteristics of items and test-takers (i.e., true difficulty, true ability level, exposure status, cheater status) to be known so accurate counts of hit rates and error rates can be made, neither of which is possible when using data from real-world testing situations. However, the simulation methodology used in this study is somewhat simplistic, which may limit the extent to which these results generalize to real-world testing situations. Aside from the influence of cheating behavior, no systematic sources of model misfit were introduced into the simulation, so the resulting data that were not manipulated were most likely unrealistically clean and unidimensional compared to data obtained from a real-world test. Other potential sources of multidimensionality, such as differential item functioning, for example, provide a possible confound for a methodology such as the one proposed in this paper.

Similar to potential concerns with the fidelity of simulation of non-manipulated item responses, it is likely that real-world manifestations of item responses obtained from exposed items are somewhat different in appearance and psychometric characteristics than their manifestation in the present study. Further research, investigating item responses submitted by individuals known to have prior exposure to test items before testing would be beneficial for improving the fidelity of cheating simulation in future studies.

Generalizing results. Performance of the novel *lco* difference method proposed in this paper was contrasted with the *lco* statistic as a means to compare the performance of this new

proposed methodology to a benchmark that is similar in application to more traditional person-fit statistics, with the goal being to draw comparisons between a method that seeks to answer the question as to whether or not the observed data fit the model (*lco*) versus a method that seeks to answer the slightly different question as to whether or not a competing model significantly improves fit (*lco* difference). The present research found evidence that there may be situations in which the proposed *lco* difference method may have more power to detect cheating when compared to more traditional approaches such as *lco*, but these findings do not necessarily generalize across all current approaches to person-fit. There are numerous, diverse person-fit statistics, which are known to have various strengths and weaknesses (Karabatsos, 2003). The results of this study indicate there may be situations in which nested model comparison has superior power when compared to the *lco* statistic, but this may not be true when comparing the proposed method to all currently-existing methods for evaluating person-fit.

Distribution of the *lco* difference. When making nested model comparisons, it is assumed that adding complexity to the model does not degrade model fit. In factor analysis, for example, adding an additional factor is not expected to negatively impact model fit indices. Adding more complexity—assuming identification issues do not come into play—is not expected to harm the fit of the model. There may be a large improvement in fit, or the improvement may be trivially small, but changes in fit are expected to always go in one direction when adding complexity. The same would be expected in the case of the *lco* difference method: the *lco* statistic computed from the one-factor model should always be larger to some extent than the *M*-*lco* statistic computed from the two-factor model (keeping in mind that poor fit is indicated by large values of these statistics and also that *lco* should have one additional degree of freedom compared to *M*-*lco*), but there were observed instances in this study where the value of

M-lco was larger than the value of *lco*, indicating paradoxically that the two-factor model had worse fit for some simulated test-takers when compared to the one-factor model. The exact reasons for this observation remain unknown at this time. It is possible that issues with rounding error associated with how output from the FA models was read and used in computation played a role, but until this issue is fully resolved, it does call into question the assumption that the *lco* difference statistic follows a χ^2 distribution. Further research is required on this topic.

Conclusion

The present study has provided evidence that comparison of changes to person-fit across nested factor analytic models may hold potential as a means to detect when cheating has occurred by means of item exposure. Additional research is warranted to compare nested model comparison to a wider range of more traditional person-fit statistics, and further work is needed to develop a satisfactory method for applying these techniques to dichotomous data. Furthermore, additional research into the sampling distribution of the proposed *lco* difference statistic is required.

References

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York: Guilford Press.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86.

Ferrando, P. J. (2007). Factor-analytic procedures for assessing response pattern scalability. *Multivariate Behavioral Research, 42,* 481-507.

Ferrando, P. J. (2009). Multidimensional factor-analysis-based procedures for assessing scalability in personality measurement. *Structural Equation Modeling, 16,* 109-133.

Han, K. T. (2007). *WinGen* [Computer software]. Amherst, MA: University of Massachusetts at Amherst.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16,* 277-298.

Meijer, R. R. & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25,* 107-135.

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. Applied Psychological Measurement, 19, 121-129.

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement, 23,* 41-53.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66,* 331-342.

Table 1

Hit Rate and Type I Error Rate Summary

Condition			*lco* Difference		*lco*	
Exposed Items	Exposed Difficulty	% Cheaters	Hit %	Type I %	Hit %	Type I %
3	Easy	1	68.1	5.4	48.9	6.8
		10	59.8	4.5	7.7	5.7
	Hard	1	81.4	5.9	87.4	6.6
		10	79.7	1.9	29.7	4.8
	Spread	1	12.5	6.1	52.0	6.9
		10	65.5	2.6	12.1	5.9
6	Easy	1	83.5	5.8	65.2	6.6
		10	78.5	2.9	44.4	4.3
	Hard	1	97.3	5.2	93.5	6.1
		10	89.0	2.4	78.3	2.5
	Spread	1	57.3	5.4	83.4	6.5
		10	86.5	2.4	52.3	4.3
	Control		-	6.1	-	7.2

Note. Hit rates provide the percentage of simulated cheaters correctly flagged by the person-fit statistic within that condition. Type I error rates for the 12 experimental conditions provide the percentage of test takers not simulated to be cheaters that were incorrectly flagged by the person-fit statistic. Type I error rates for the control condition provide the percentage of simulated test-takers flagged by the person-fit statistics in the condition in which no cheating was simulated.

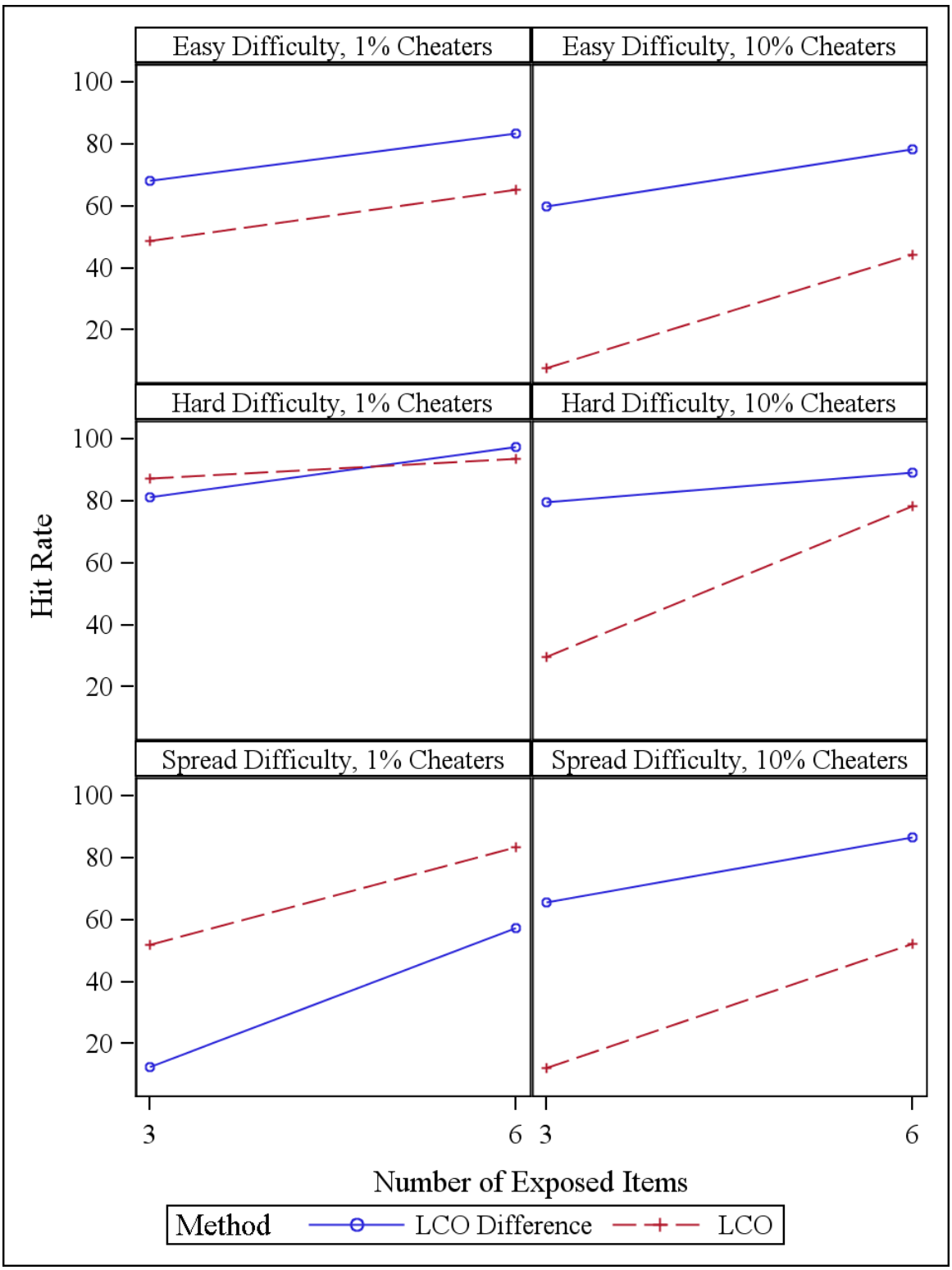

Figure 1. Hit rate comparison across experimental conditions.