# Effect of Test Order When Administering Multiple Rating Scales to a Single Rater

Janice Kooken[1], Megan E. Welsh[1], Faith G. Miller[1], T. Chris Riley-Tillman[2], & Sandra M. Chafouleas[1]

*University of Connecticut[1] & University of Missouri[2]*

## Introduction

**Background**

The use of counterbalancing in scientific research is well established regarding treatment order as a method of reducing threats to internal validity. In educational and psychological research, the need for counterbalancing self-report measures has been examined (Cunningham, Preacher & Banaji, 2001), but assessment presentation order is often overlooked and rarely incorporated into the design of studies utilizing raters. As a result, the effect of order of multiple test administrations to a single rater in behavior protocols has not been thoroughly examined (Lucas, 1992).

**Objective**

The current study presents the results of an analysis of order effects utilizing data from a larger validation study of the Direct Behavior Rating Single Item Scales (DBR-SIS).

The study examined the following questions:

1. Are there significant differences in the multivariate measure of behavior due to changes in test order?
2. After controlling for behavior within classrooms, does test order results in significant differences in behavior ratings?

## Method

**Participants and Setting**

For this study, a total of 1976 students were rated by teachers on all three measures in the fall, winter and spring. Students were mostly Caucasian (82%), with 7% Hispanic, and 13% students with disabilities. Students were nested within teachers, with 202 teachers participating in 23 urban, suburban, and rural school districts in 3 states: Connecticut, New York, and Missouri.

**Measures**

*The Direct Behavior Rating – Single Item Scale* (DBR-SIS) reflects the teacher's perception of the proportion of time a student is observed engaged in a target behavior (academic engagement, respectful, disruptive) from 0 (never) to 10 (always). Students were rated twice daily for five days. The composite rating is the sum of the AE, RS, and 10-DB scores. The *Social Skills Improvement System - Performance Screening Guide* (SSiS; Gresham & Elliott, 1990) can be used to screen social and academic behaviors of all students in a class. This measure is comprised of four scales, but only the Prosocial Behavior scale was used in this study. The *Behavioral and Emotional Screening System* (BESS; Kamphaus & Reynolds, 2007) is a brief rating scale that can be useful in screening for behavioral and emotional strengths and weaknesses in children and adolescents.

A composite DBR-SIS score was computed using a sum of the individual scores, resulting in a minimum score of 0 and maximum of 30. This composite was averaged over the two ratings completed on each day. Since the DBR-SIS single item scales lack characteristics of a normal distribution, the composite was then squared to improve normality.

## Method

**Procedures**

Of the six possible ways to sequence the three measures, one was assigned at random to each teacher, distributed equally across grades and sites.\

Table 1. *Test order within condition*

| Condition | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| First | DBR | DBR | BESS | BESS | SSIS | SSIS |
| Second | BESS | SSIS | DBR | SSIS | DBR | BESS |
| Third | SSIS | BESS | SSIS | DBR | BESS | DBR |

Two different techniques were utilized to identify order effects.

1. A MANOVA tested whether there exists a linear combination of the behavior rating variables for which significant differences exist by condition, site, and condition by site. The DBR-SIS was found to be heteroscedastic and leptokurtic, so a nonparametric method using ranked data and Pillai's trace was employed (Finch, 2005).
2. To control for the nested nature of the data, with students nested within teachers, hierarchical linear models measured the effect of test order on the dependent variable $Y_{ij}$, consisting of either the composite DBR-SIS, the SSIS Prosocial scale, or the BESS combined T score, while controlling for student behavior characteristics within classroom. The level 1 and level 2 equations are as follows:

**Level-1 Model**

$$Y_{ij} = \beta_{0j} + \beta_{1j}*(GRADE_{ij}) + \beta_{2j}*(BESST_{ij}) + \beta_{3j}*(SSISPROS_{ij}) + r_{ij}$$

**Level-2 Model**

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Order_j) + u_{0j}$$
$$\beta_{1j} = \gamma_{01}$$
$$\beta_{2j} = \gamma_{20}$$
$$\beta_{3j} = \gamma_{30}$$

Variables were created to reflect the order of administration of other tests relative to the dependent variable. Eighty-one models were run to test effects of the order effect variables on Lower elementary (grade 1 &2), upper elementary (grade 4&5), and middle school (grade 7&8) separately for fall, winter and spring.

## Results

1. Results of the MANOVA indicated statistically significant order effects for all three data collections, by condition, site, and condition crossed and site.

Table 2: *Results of MANOVA.*

| | Fall | | Winter | | Spring | |
|---|---|---|---|---|---|---|
| Effect | p | Effect Size (partial $\eta^2$) | p | Effect Size (partial $\eta^2$) | p | Effect Size (partial $\eta^2$) |
| Condition | <.001 | 0.013 | <.001 | 0.016 | <.001 | 0.015 |
| Site | <.001 | 0.03 | <.001 | 0.032 | <.001 | 0.029 |
| Condition by Site | <.001 | 0.018 | <.001 | 0.017 | 0.004 | 0.013 |

## Results

2. Of the eighty-one models, only four combinations of order effects were found to have a significant impact on ratings as shown below in Table 3.

Table 3. *Effect of test order in Multilevel model*

| | | | Fall | | | Winter | | | Spring | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | DV | Order tested | LE | UE | MS | LE | UE | MS | LE | UE | MS |
| 1 | DBR | DF | | | | | | | | | |
| | DBR | C | | | | | | | | | |
| | DBR | E | | | | | | | | | |
| 2 | DBR | CDF | | | | | | | | | |
| 3 | DBR | DEF | | | | | | | | | |
| 4 | SSIS | B | | | | | | | | | |
| | SSIS | AC | | | | | | | | | |
| | SSIS | D | | | | | | | | | |
| 5 | SSIS | ACD | | | | | | | | | |
| 6 | SSIS | ABC | | | | | | | | | |
| 7 | BESS | BE | | | | | | | | | |
| | BESS | F | | | | | | | | | |
| | BESS | A | | | | | | | | | |
| 8 | BESS | BEF | | | | | | | | | |
| 9 | BESS | ABE | | | | | | | | | |

## Summary and Conclusions

While test order was found to be associated with teacher's global perceptions of students in the MANOVA analyses, these results are somewhat mitigated by the lack of random assignment of students to classrooms. Therefore, it is possible that the characteristics of students within clusters impact these results. The multilevel models, which do control for student behavior, only examine one behavioral measure at a time and seem to indicate that test order does not impact teacher ratings.

**Limitations and Implications for Future Research**

In the current study, since students were nested within teachers, the impact of teacher characteristics was confounded with test order effects. In addition, without a universally established behavior measure, each of the rating scales was used alternatively as the predicted and control variables, further confounding results. Further analysis is needed, including, (1) changing the distributional assumptions associated with the analyses to better account for lack of normality in DBR-SIS composites and (2) examining the impact of order effects while controlling for student-level behavior using measures (and perhaps raters) unrelated to the measures of interest.

Preparation of this poster was supported by a grant from the Institute for Education Sciences (IES), U.S. Department of Education (R324B060014). For additional information, please direct all correspondence to Sandra Chafouleas at sandra.chafouleas@uconn.edu