PSYCHOMETRIKA—VOL. 83, NO. 4, 941–962
DECEMBER 2018
https://doi.org/10.1007/s11336-018-9631-z

PSYCHOMETRIC MODELS OF SMALL GROUP COLLABORATIONS

PETER F. HALPIN AND YOAV BERGNER
NEW YORK UNIVERSITY

The social combination theory of group problem solving is used to extend existing psychometric models to collaborative settings. A model for pairwise group work is proposed, the implications of the model for assessment design are considered, and its estimation is addressed. The results are illustrated with an empirical example in which dyads work together on a twelfth-grade level mathematics assessment. In conclusion, attention is given to avenues of research that seem most fruitful for advancing current initiatives concerning the assessment of collaboration, teamwork, and related constructs.

Key words: item response theory, social combination theory, process loss, synergy, group work.

1. Introduction

Webb (1995) discussed three main purposes of small group collaborations in assessment settings: (1) to measure group performance, (2) to measure individual performance following group collaboration, and (3) to measure individuals' teamwork skills. This paper focuses on the first of these purposes. In particular, it is argued that the social combination approach to group problem solving (see, e.g., Laughlin, 2013, chap. 2) provides a suitable framework for generalizing existing psychometric models to small group settings. The need for such models is underscored by recent initiatives concerning the assessment of collaborative problem solving, teamwork, and related constructs (e.g., Heckman & Kautz, 2014; Herman & Hilton, 2017; Fiore et al., 2017; Griffin & Care, 2015; Lippman et al., 2015; National Research Council, 2011, 2015; OECD, 2017; Pellegrino & Hilton, 2012; von Davier et al., 2017).

The literature on social combination theory is selectively reviewed to motivate a general modeling framework. In Sect. 2, we use the general framework to specify a model of pairwise group work. In Sect. 3, the utility of the proposed model is illustrated by deriving a number of results about the design of group assessments. In particular, we consider how to design group assessments that can provide empirical evidence of *synergy*. Larson (2010) defined (strong) synergy as occurring when the performance of a group exceeds that expected of its most capable member working individually, and he documented the relative lack of research on this topic. Section 4 addresses parameter estimation. The results are illustrated with data simulation (Sect. 5) and an empirical example in which pairs of respondents work together to complete a twelfth-grade level mathematics assessment (Sect. 6).

This paper focuses on binary (correct/incorrect) response data, unidimensional latent traits for individual performance, and groups that are independent from one other (i.e., the groups do not share members). Additionally, due to the nature of the example data, this research has so far been limited to inferential methods that can be applied when only a relatively small number of dyads are observed working together (e.g., we do not estimate item parameters under group testing conditions). In the concluding section, these limitations are discussed in terms of specific lines of future research that seem fruitful for advancing current initiatives concerning the assessment of collaboration, teamwork, and related domains.

Correspondence should be made to Peter F. Halpin, New York University, 246 Greene Street, Office 204, New York, NY10002, USA. Email: peter.halpin@nyu.edu

1.1. Overview of Social Combination Theory

1.1.1. Task Types It is useful to begin by delineating the types of tasks that are under consideration. McGrath's (1984) circumplex typology distinguishes tasks according to two theoretical dimensions. One dimension represents a contrast between tasks that incentivize either cooperation or conflict among group members. The second dimension represents a contrast between tasks that are principally cognitive versus behavioral in nature. *Intellective* tasks (Laughlin, 1980) are located in the cooperative-cognitive quadrant of the circumplex and are characterized by problem-solving scenarios in which there exists a demonstrably correct answer. This type of task is exemplified by problems in mathematics and logic, as well as problems that are about factual content. By contrast, *decision-making* tasks are problem-solving scenarios in which no agreed-upon correct answer exists. This type of task is exemplified by jury deliberation, and it typically involves some degree of conflict among group members (e.g., Davis, 1992). In the literature on cooperative learning, a similar distinction has been made between well-structured and ill-structured group activities (Cohen, 1994). Although the statistical models used in social combination theory are applicable to any task in which the outcome can be represented as a categorical variable, the present research focuses on tasks that require individuals to cooperate to provide demonstrably correct responses.

1.1.2. Defining Group Responses What constitutes a correct response in a group setting? To answer this question, it is useful to consider the distinction between *unitary* and *divisible* tasks (Steiner, 1972). A unitary task requires a single result or product from all group members, whereas a divisible task permits different outputs from different subsets of group members. In an assessment context, this distinction can be interpreted in terms of how a task is scored. For example, we may wish to score the group as a whole or the responses of individual members separately. While these objectives need not be incompatible, the focus of the present research is to evaluate the group as a whole.

The notion of a group response can be formalized in terms of scoring rules that are applied to the responses of individual members. The types of unitary tasks identified by Steiner (1972) provide a number of plausible scoring rules. For example, when using a disjunctive scoring rule, a group is regarded as producing a correct response only if any of its members do. Conversely, a conjunctive scoring rule defines a group's response as correct only if all of its members provide a correct response. A number of psychometric models make similar distinctions, such as compensatory versus non-compensatory multidimensional item response theory models (e.g., Reckase, 2009), and "noisy-and" versus "noisy-or" models in cognitive diagnostic assessment (e.g., Junker & Sijtsma, 2001). As an intuitive starting point, this research focuses on conjunctive scoring rules.

1.1.3. Modeling Group Performance A key concept of social combination theory is the *decision function*, which was introduced in the following quotation from Smoke and Zajonc (1962):

> If p is the probability that a given individual member is correct, the group has a probability $h(p)$ of being correct, where $h(p)$ is a function of p depending upon the type of decision scheme accepted by the group. We shall call $h(p)$ a decision function. Intuitively, it would seem that a decision function is desirable to the extent that it surpasses p (p.322).

This quotation summarizes two seminal ideas. First, that group performance can be modeled as a function of individual performance. Second, that group performance can be evaluated by comparing among alternative models. Social combination research has largely focused on testing theoretically motivated decision functions against data aggregated over groups (see Larson, 2010;

Laughlin, 2013). By contrast, the present research focuses on the estimation of parametric decision functions that characterize differences among groups.

2. Model Specification

The basic ingredients of social combination theory are: (a) a model for individual performance and (b) a function for mapping individual performance to group performance. We provide (a) in terms of an item response theory (IRT) model for binary data. We provide (b) by first specifying group assessments such that they include the IRT model as the special case where all groups have a single member. We then use Davis' (1973) social combination model to generalize to cases where groups have multiple members. Finally, we propose a model for pairwise group performance and discuss its interpretation.

2.1. Individual Assessments

Let X_{ij} denote a Bernoulli random variable representing whether respondent $j = 1, \ldots, J$ answers item $i = 1, \ldots, I$ of a test T correctly ($X_{ij} = 1$) or incorrectly ($X_{ij} = 0$). In this paper we stipulate that the response vector $\boldsymbol{X}_j = (X_{1j}, \ldots, X_{Ij})$ can be described in terms of a univariate monotone latent variable (UMLV) model (Holland & Rosenbaum, 1986, sections 2.1–2.3). Assuming the existence of a latent variable $\theta \in \mathbb{R}$, a UMLV model is defined via the following two conditions on the joint distribution of $\boldsymbol{X}_j$ and θ_j. First, the item responses $\boldsymbol{X}_j$ are conditionally independent given θ_j:

$$F(\boldsymbol{X}_j \mid \theta_j) = \prod_{i=1}^{I} F_i(X_{ij} \mid \theta_j). \tag{1}$$

The second requirement is latent monotonicity of the item response functions (IRFs), which can be stated in terms of the expected value of X_{ij} given θ_j:

$$E(X_{ij} \mid \theta_j) \leq E(X_{ij} \mid \theta_j'), \tag{2}$$

for $\theta_j < \theta_j'$. We let $P_{ij} = P_i(\theta_j) = E(X_{ij} \mid \theta_j)$ and $Q_{ij} = 1 - P_{ij}$. We also make explicit the assumption of independence of respondents, which is required in Sect. 2.3. Letting $\mathcal{X} = \{X_{ij}\}$ denote the response matrix and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$, this assumption is

$$F(\mathcal{X} \mid \boldsymbol{\theta}) = \prod_{j=1}^{J} F(\boldsymbol{X}_j \mid \theta_j). \tag{3}$$

2.2. Group Assessments

Consider the case where respondents are assigned without replacement to groups. Let the set $\mathcal{J} = \{j[1], \ldots, j[J]\}$ denote the respondents and G denote a partition of $\mathcal{J}$ with elements $G_k = \{j[k_1], \ldots, j[k_n]\}$ that satisfy

$$G_k \cap G_{k'} = \emptyset, \quad k \neq k', \quad \text{and} \quad \cup_{k=1}^{K} G_k = \mathcal{J}.$$

Then n is the number of respondents in group k, here assumed to be constant, and $K = J/n$ is the number of non-overlapping groups formed from a pool of J respondents.

Let Y_{ik} denote a Bernoulli random variable representing whether group $k = 1, \ldots, K$ answers item $i = 1, \ldots, I$ of a test T' correctly ($Y_{ik} = 1$) or incorrectly ($Y_{ik} = 0$). As mentioned, the focus of the present paper is group responses that result from applying a conjunctive scoring rule to the responses of individual group members, say $Y_{ik} = \prod_{r \in G_k} Y^*_{ir}$, with Y^*_{ir} denoting the individual responses.

Similar to the UMLV model for individual assessments, we will be interested in models where the response vector $\boldsymbol{Y}_k = (Y_{1k}, \ldots, Y_{Ik})$ is conditionally independent given a vector of group parameters $\boldsymbol{\zeta}_k$ that includes $\boldsymbol{\theta}_k = (\theta_{k_1}, \ldots, \theta_{k_n})$, as well as additional parameters introduced below to characterize the group decision function. Formally,

$$F(\boldsymbol{Y}_k \mid \boldsymbol{\zeta}_k) = \prod_{i=1}^{I} F_i(Y_{ik} \mid \boldsymbol{\zeta}_k), \tag{4}$$

and we let $R_{ik} = R_i(\boldsymbol{\zeta}_k) = E(Y_{ik} \mid \boldsymbol{\zeta}_k)$ denote the group IRFs. We also assume that the responses of the different groups are independent,

$$F(\mathcal{Y} \mid \mathcal{Z}) = \prod_{k=1}^{K} F(\boldsymbol{Y}_k \mid \boldsymbol{\zeta}_k), \tag{5}$$

where $\mathcal{Y} = \{Y_{ik}\}$ is the response matrix and $\mathcal{Z} = \{\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_K\}$.

When $n = 1$, the definition of a group responses reduces to that of an individual response. In this case, we additionally stipulate that $R_i(\boldsymbol{\zeta}_k) = P_i(\theta_j)$ for $j \in G_k$—i.e., when there is only one person in each group, the group IRF is just an individual IRF, as defined in Eq. (2). In more practical terms, we are assuming that each group assessment has a corresponding individual assessment that is identical other than the instructions pertaining to how respondents may work together. We refer to this as the *individual version of a group assessment*. When $n > 1$, the goal is to derive the properties of $R_i(\boldsymbol{\zeta}_k)$ using the $P_i(\theta_j)$, $j \in G_k$, and the following model for group performance.

2.3. A General Social Combination Model

Davis (1973) presented a general formulation of social combination models that we adapt to the present context as follows. Let $\boldsymbol{X}^*_{ik} = (X^*_{ik_1}, \ldots, X^*_{ik_n})$ denote the responses that would have resulted if each member of group k had written the individual version of item i on a group assessment T'. There are $S = 2^n$ possible realizations of $\boldsymbol{X}^*_{ik}$, each with probability

$$\pi_{iks} = \text{Prob}(\boldsymbol{X}^*_{ik} = x_s \mid \boldsymbol{\theta}_n) = \prod_{r=1}^{n} P_{ik_r}^{x_{ik_r}} Q_{ik_r}^{1-x_{ik_r}}, \quad s = 1, \ldots, S \tag{6}$$

following directly from Eq. (3). Note that π_{iks} is the probability of n responses to a single item i, not to be mistaken with the more familiar expression for I responses of a single individual.

For a binary group response, a social combination model can be specified as $2 \times S$ matrix $D_k = \{d_{rs}\}$ that maps the probabilities of the individual responses, $\boldsymbol{\pi}_{ik} = (\pi_{ik1}, \ldots, \pi_{ikS})$, onto the probabilities of the group responses $\boldsymbol{\rho}_{ik} = (R_{ik}, 1 - R_{ik})$. The general model to be considered is then

$$\boldsymbol{\rho}_{ik} = D_k \boldsymbol{\pi}_{ik}, \tag{7}$$

in which it is required that $d_{rs} \in [0, 1]$, and $\sum_r d_{rs} = 1$ for each column $s = 1, \ldots, S$, to ensure that $R_{ik} \in [0, 1]$.

In comparison with the model considered by Davis (1973), Eq. (7) (a) does not assume that the probability of a correct response to an item is equal for all individuals, (b) allows these probabilities to vary over items, and (c) allows the decision matrix D_k to vary over groups. It may therefore be interpreted as a psychometric reformulation of social combination theory.

2.4. A Restricted Model for Pairwise Group Performance

In the remainder of this paper we confine attention to a special case of Eq. (7) that arises for groups of size $n = 2$. We focus on a structurally constrained parameterization of the decision matrix, which we refer to as the restricted social combination (RSC) model for dyads:

$$\begin{bmatrix} R_{ik} \\ 1 - R_{ik} \end{bmatrix} = \begin{bmatrix} 1 & a_k & b_k & 0 \\ 0 & 1 - a_k & 1 - b_k & 1 \end{bmatrix} \times \begin{bmatrix} P_{ik_1} P_{ik_2} \\ P_{ik_1} Q_{ik_2} \\ Q_{ik_1} P_{ik_2} \\ Q_{ik_1} Q_{ik_2} \end{bmatrix}, \tag{8}$$

where $a_k, b_k \in [0, 1]$. The corresponding group IRF is

$$R_{ik} = P_{ik_1} P_{ik_2} + a_k P_{ik_1} Q_{ik_2} + b_k Q_{ik_1} P_{ik_2}, \tag{9}$$

and the vector of group parameters is $\boldsymbol{\zeta}_k = (a_k, b_k, \theta_{k_1}, \theta_{k_2})$. Going forward, we simplify notation by omitting the group index k and writing $k_r = r$ to denote the members $r = 1, 2$ of an arbitrary dyad.

The decision parameters a and b govern the probability of a correct group response in cases where only one of the partners would have provided a correct response when working individually. As discussed below, identification of these parameters can be problematic for some combinations of θ_1 and θ_2, leading us to consider a "one-parameter" version of the model with $a = b = w \in [0, 1]$ for data analytic purposes. However, before moving on to parameter estimation, we consider the properties of the model in Eq. (8). The remainder of this section outlines the interpretation of the structural restrictions placed on the decision matrix. The following section presents a number of results about assessment design, which are of interest in their own right and also lead to conditions for identifying the decision parameters (see Proposition 7).

2.4.1. Motivation for Parameter Restrictions The RSC model is restricted in the sense that the first and last columns of the decision matrix contain structural zeros. Intuitively, these restrictions imply the omission of parameters for "slipping" (first column) and "guessing" (last column). Since we have required that $R_{ik} = P_{ij}$ when $n = 1$, it is natural to omit these parameters when extending the model to cases where $n > 1$. Additionally, when a guessing parameter is included, there is no guarantee that the group IRF will be latent monotonic. On the other hand, the following proposition shows that the RSC model preserves latent monotonicity (see "Appendix" for proof).

Proposition 1. *Let $R_i(\boldsymbol{\theta})$ denote a group IRF in Eq. (9), treated as a function $\boldsymbol{\theta} \in \mathbb{R}^2$, with parameters $a, b \in [0, 1]$. If $P_i(\theta)$ satisfies latent monotonicity [see Eq. (2)], then $R_i(\boldsymbol{\theta})$ also satisfies latent monotonicity, in each coordinate of $\boldsymbol{\theta}$.*

TABLE 1.
Four special cases of the RSC model.

a	b	Group IRF	Model name	Interpretation
0	0	$R_i^{\text{Ind}} = P_{i1}\, P_{i2}$	Independence	Members do not work together (lower bound)
1	0	$R_i^{\text{Min}} = P_{i1}$	Individual (Min)	The less able member performs the task
0	1	$R_i^{\text{Max}} = P_{i2}$	Individual (Max)	The more able member performs the task
1	1	$R_i^{\text{Add}} = 1 - Q_{i1}\, Q_{i2}$	Additive	One example of group synergy (upper bound)

The interpretations assume that a conjunctive scoring rule is used and that $\theta_1 \leq \theta_2$.

The structural restrictions in the decision matrix also impose non-trivial lower and upper bounds on R_i. The interpretation of the bounds, as well as two other special cases, is summarized in Table 1. The independence model describes the case where group members do not work together. Given appropriate testing instructions, we expect that the independence model would be a reasonable lower bound on empirical group performance. For example, respondents might be instructed that they may choose to work without their partner at any point during the test.

The RSC model also includes as special cases the individual performance of either group member. As described by Webb (1995), various kinds of participation biases can lead to this type of situation. For example, it can be both efficient and effective for groups to defer output to the most able individual, leading to the maximum individual ("Max") model. On the other hand, the minimum individual ("Min") model may arise when group members' participation is influenced by status characteristics that are not related to their ability.

The Additive model is a psychometric reformulation of Lorge and Solomon's (1955) model A. Its justification as an upper bound is motivated by a large number of experimental studies showing that group performance on intellective tasks very rarely exceeds the level predicted by this model (see reviews by Davis, 1992; Larson, 2010; McGrath, 1984; Steiner, 1972). Steiner (1972) also provided a theoretical rationale, based on information sharing, that supported the interpretation of Model A as the ideal or maximal group performance on intellective tasks. On this interpretation, values of $a, b < 1$ correspond to what Steiner termed *process loss*, which describes the discrepancy between a group's theoretical maximum performance and its actual performance.

In summary, the RSC model for dyads preserves latent monotonicity, places lower and upper bounds on group performance that are reasonable from the perspective of assessment design and past research, and includes as special cases the IRT models describing the performance of either individual.

3. Assessment Design

In this section we illustrate the utility of the RSC model by deriving a number of results about the design of group assessments using the four special cases outlined in Table 1. These results provide insights about how to select items and match partners so that the four models imply distinct joint distributions for the group response vector $\boldsymbol{Y}$. As mentioned in introduction, we are especially concerned to design assessments that can provide evidence about group synergy, so we frame the analysis around a formal definition of this concept. The final result links the discussion of assessment design to parameter estimation by deriving conditions for identifying the decision parameters.

3.1. Preliminaries

It is readily verified that the four group IRFs considered in Table 1 are related as follows:

$$R_i^{\text{Ind}} \leq R_i^{\text{Min}} \leq R_i^{\text{Max}} \leq R_i^{\text{Add}}. \tag{10}$$

The reader may have already noted a number of conditions under which these inequalities can be replaced by strict equalities. Most obviously, if $\theta_1 = \theta_2$, then $R_i^{\text{Min}} = R_i^{\text{Max}}$. Using the terminology of Vuong (1989, Def. 3), the Min model and the Max model are *overlapping*, because they imply the same distribution of $\boldsymbol{Y}$ for some values of $\boldsymbol{\theta}$, but neither model is nested within the other. We refer to situations in which models overlap as a problem of model equivalence and consider how to design group assessments so as to avoid equivalence among the four models.

Requiring that the Additive and Max models are non-equivalent yields the condition:

$$\Delta_i(\boldsymbol{\theta}) \equiv R_i^{\text{Add}}(\boldsymbol{\theta}) - R_i^{\text{Max}}(\boldsymbol{\theta}) = P_i(\theta_1)\, Q_i(\theta_2) > 0 \tag{11}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and $\theta_1 \leq \theta_2$. We make use of the simplified notation $\Delta_i = \Delta_{i12} = \Delta_i(\boldsymbol{\theta})$, omitting the subscripts for persons where this is clear from context.

Larson's (2010) definition of (strong) group synergy requires that observed group performance exceeds R_i^{Max}. Consequently, the requirement that $\Delta_i > 0$ ensures that a group assessment based on the RSC model can provide evidence of synergy. To this end, we derive a number of results that characterize Δ_i as a function of item selection (i.e., properties of P_i) and team composition (i.e., the values of θ_1 and θ_2).

Analysis of Δ_i is also sufficient to describe equivalence between the Min and Individual models, which follows from noting that $\Delta_i = R_i^{\text{Min}} - R_i^{\text{Ind}}$. There is relatively little literature addressing worse-than-individual group performance (i.e., cases where observed group performance is worse than R_i^{Min}; see Mathieu et al., 2008). We will refer to this situation as group *antagonism*. It is a convenient feature of the RSC model that synergy and antagonism can be addressed by analysis of the same quantity, and in particular that an assessment designed to ensure $\Delta_i > 0$ will be useful for providing evidence about both types of group performance.

Concerning the Min and Max models, the problem of model equivalence is addressed by well-known results in IRT. In particular, $R_i^{\text{Max}} - R_i^{\text{Min}} = P_{i2} - P_{i1}$ is monotone non-decreasing in $\delta = \theta_2 - \theta_1$ and will be larger for highly discriminating items. As we now show, these are in fact the conditions under which $\Delta_i = 0$.

3.2. Analytic Results About the Design of Group Assessments

First we note that $\Delta_i = 0$ if and only if $P_{i1} = 0$ or $P_{i2} = 1$. This follows trivially from the requirement $\theta_1 \leq \theta_2$ and latent monotonicity of P_i [see Eq. (2)]. These two situations correspond to what Shiflett (1979) described as *redundancy*, because the contributions of the less able partner cannot add anything to performance of the more able partner. We therefore refer to $\Delta_i = 0$ in terms of redundancy of items or of team members. To avoid confusion with established uses of the term "information" in IRT, we do not use it as an antonym for redundancy.

Proposition 2 addresses how to avoid redundancy when designing group assessments. The proofs for all propositions are contained in "Appendix."

Proposition 2. *If* $0 < P_{i1} \leq P_{i2} < 1$, *then* $\Delta_i(P_{i1}, P_{i2}) = P_{i1}Q_{i2}$ *is strictly concave, with global maximum* $\Delta_i(1/2, 1/2) = 1/4$.

The following result describes a special case of Proposition 2 that applies to many IRT models of binary data.

Proposition 3. *If $P_i(\theta)$ is strictly increasing on a neighborhood $\mathcal{N}$ around $\theta_i^* = \{\theta \mid P_i(\theta) = .5\}$, then*

Part 1 $\arg\max_{\boldsymbol{\theta}} \{\Delta_i(\boldsymbol{\theta})\} = (\theta_i^*, \theta_i^*)$.

Part 2 For $\theta_1 \leq \theta_i^* \leq \theta_2 \in \mathcal{N}$, $\Delta_i(\boldsymbol{\theta})$ is strictly decreasing with $\delta = \theta_2 - \theta_1$.

Part 1 of Proposition 3 states that redundancy will be minimized when both group members have ability equal to the difficulty of the item. Part 2 states that redundancy is strictly increasing as the ability level of either partner moves away from the difficulty level of the item.

In practice, it generally will not be feasible to select partners and items to satisfy Part 1 of Proposition 3, especially when the pool of examinees and/or items is quite small. To address this situation, the next three propositions describe the problem of item selection for cases where $\delta \geq 0$. However, all of the following results also require stronger assumptions on the UMLV model for individual performance, which are stated as part of the following proposition.

Proposition 4. *Assume that the IRFs for individual performance can be written as a two-parameter logistic (2PL) model*

$$P_i(\theta) = [1 + \exp\{-\alpha_i(\theta - \beta_i)\}]^{-1} \tag{12}$$

with $\beta_i \in \mathbb{R}$ and $\alpha_i > 0$. Let $\Delta_i(\beta_i)$ denote the item delta in Eq. (11), treated as a function of β_i for any fixed values of $\theta_1 \leq \theta_2$. Then

$$\beta_i^* \equiv \arg\max_{\beta_i} \{\Delta_i(\beta_i)\} = (\theta_1 + \theta_2)/2$$

and $\Delta_i(u)$ is monotone decreasing in $u = |\beta_i - \beta_i^*|$.

Although the assumption of a 2PL IRF is quite restrictive, the result is nonetheless interesting because of its simplicity. For any pair of respondents, items chosen to have difficulty equal to the average of the group members' abilities will be least redundant, and items will become increasingly redundant the farther they move away from this optimal value.

The next proposition states that the minimum item redundancy is increasing with the discrimination of the item, implying that highly discriminating items will not, in general, be well suited for measuring group synergy.

Proposition 5. *Assume that the IRFs for individual performance can be written as in Proposition 4 and let $\Delta_i^*(\alpha_i)$ denote the item delta in Eq. (11) treated as a function of α_i, evaluated at β_i^*, for any fixed values of $\theta_1 \leq \theta_2$. Then for $\alpha_i < \alpha_i'$,*

$$\Delta_i^*(\alpha_i) \geq \Delta_i^*(\alpha_i').$$

When $\theta_1 = \theta_2$, $\Delta_i^*(\alpha_i) = 1/4$ for any value of $\alpha_i > 0$. Otherwise, $\Delta_i^*(\alpha_i) \to 0$ as $\alpha_i \to \infty$.

The following result characterizes a "trade-off" between team composition and item selection, again assuming a 2PL IRF.

Proposition 6. *Under the same conditions as Proposition 5, let $\Delta_i^*(\alpha_i) = D$ for some constant $D \in (0, 1/4]$. Then*

$$\alpha_i = \frac{2}{\delta} \ln \frac{1 - \sqrt{D}}{\sqrt{D}}.$$

The proposition shows that, for any desired level of item (non-) redundancy, $D > 0$, the item discrimination must be chosen to be inversely proportional to the difference between the ability levels of the respondents.

Finally, we relate these results on model equivalence to identification of the parameters of the decision matrix.

Proposition 7. *When $\Delta_i \in \{0, 1/4\}$, the decision function parameters of the RSC model presented in Eq. (8) are not identified. Letting R_i denote the group IRF in Eq. (9),*

Part 1 If $\Delta_i = 0$, then $E\left[\frac{\partial^2}{\partial a^2} \ln f(Y_i \mid \boldsymbol{\zeta})\right] = \frac{\Delta_i^2}{R_i(1-R_i)} = 0$.

Part 2 If $\Delta_i = 1/4$, then $R_i(\boldsymbol{\zeta}) = R_i(\boldsymbol{\xi})$ where $\boldsymbol{\zeta} = (\alpha, \beta, \theta_1, \theta_2)$ and $\boldsymbol{\xi} = (\beta, \alpha, \theta_1, \theta_2)$.

Part 1 shows that the item information (the expected Fisher information) of the parameter a is equal to zero when $\Delta_i = 0$. Part 2 shows that the exchanging the decision parameters a and b implies the same group IRF and consequently the same model-implied distribution of Y_i, when $\Delta_i = 1/4$.

3.3. Summary

The practical implications of this section are as follows: an assessment designed to provide evidence of group synergy should (a) match group members with similar levels of ability (Propositions 2 and 3), (b) select items that are targeted between their ability levels (Proposition 4), and (c) avoid the use of highly discriminating items (Propositions 5). In particular, Proposition 6 showed that it will not be possible to select items that are both highly discriminating (i.e., strongly related to the performance domain) and also non-redundant (i.e., provide evidence of synergy), when group members have highly disparate levels of ability. These results are all predicated on the RSC model for dyads presented in Eq. (8), and Propositions 4, 5, and 6 additionally assumed a 2PL model for individual performance.

Proposition 7 summarizes the implications of these results for estimating the decision parameters of the RSC model. When items are entirely redundant (i.e., $\Delta_i = 0$), they provide no information about the parameter a. On the other hand, when items minimize redundancy (i.e., $\Delta_i = 1/4$), the individual parameters are not unique (although their sum is). Referring to Proposition 3, we may interpret these results in terms of team composition: the decision parameters are not identified when groups members' abilities are either "too proximate" ($\delta \to 0$) or "too disparate" ($\delta \to \infty$). In practice we have found that parameter recovery can be quite unstable for a relatively wide range of values of δ, leading us to consider alternative approaches to parameter estimation.

4. Parameter Estimation

4.1. The One-Parameter RSC Model

To avoid the cases of model un-identification described in Proposition 7, we propose to address data analysis and inference using a "one-parameter" RSC model for dyads obtained by setting $a = b = w \in [0, 1]$:

$$R_i^* = w(P_{i1}\, Q_{i2} + Q_{i1}\, P_{i2}) + P_{i1}\, P_{i2}. \tag{13}$$

This parameterization of the model has the following three desirable properties.

1. For each item i, the observed Fisher information of the weight w (see Eq. (21) of "Appendix") is strictly greater than zero unless $P_{ir} \in \{0, 1\}$ for both $r = 1$ and $r = 2$. Hence, w is identified under the same conditions as θ_1 and θ_2.
2. The group IRF may be written as a linear interpolation between the lower and upper bounds of the RSC model,

$$R_i^* = w\, R_i^{\text{Add}} + (1 - w)\, R_i^{\text{Ind}}. \tag{14}$$

   The parameter w is then directly interpretable in terms of Steiner's (1972) concept of process loss, with $w = 0$ denoting complete process loss, and $w = 1$ denoting maximal group performance.
3. Setting $w = 1/2$ implies $R_i^* = (P_{i1} + P_{i2})/2$, which facilities the interpretation of w in terms of group synergy. In particular, when group members are matched on ability, $w = 1/2$ describes their expected level of performance when working individually, and consequently $w > 1/2$ corresponds to strong synergy. When group members are not matched on ability, $w > 1/2$ corresponds to what Larson (2010) referred to as weak synergy (i.e., better performance than the average group member working individually). Corresponding considerations apply to group antagonism.

In summary, the one-parameter RSC model in Eq. (14) is identified under more general conditions than its two-parameter counterpart in Eq. (9). The weight parameter w can be interpreted in terms of process loss and can also be used to evaluate group performance with respect to synergy and antagonism.

4.2. Estimating the One-Parameter RSC Model

We use data from both an individual assessment and a group assessment to estimate the group parameter vector of the one-parameter RSC model, $\boldsymbol{\zeta}^* = (v, \theta_1, \theta_2)$, with $v = \text{logit}(w)$. The logit parameterization facilitates estimation when w approaches the boundary values $\{0, 1\}$. "Appendix" contains the maximum likelihood (ML) equations for estimating $\boldsymbol{\zeta}^*$.

We also provide equations for modal a' posteriori (MAP) estimation, which improves parameter recovery when relatively few items are available on the individual or group assessments (see Sect. 5). For θ_1 and θ_2 we use a standard normal prior; for v we use a weakly informative normal prior centered at zero. The rationale for centering at $\mu_v = 0$ is given by point 3 of the previous section. We use $\sigma_v = 3$ to define "weakly informative" in the present context. Then $\mu_v \pm \sigma_v$ corresponds to $w \in [.05, .95]$ and $\mu_v \pm 2\sigma_v$ corresponds to $w \in [.002, .998]$. We do not expect to be able to precisely recover values of v outside of this range.

In addition to ML and MAP we have also implemented Bayesian estimation of $\boldsymbol{\zeta}^*$ using the `Stan` language; results with simulated data indicated that performance is very similar to that of the MAP estimator using observed Fisher information for the standard errors. Expected a' posterior in three dimensions also remains tractable by numerical integration. A "fully" Bayesian approach could be used to incorporate sampling uncertainty in the item parameters. Plausible values might be used when item-level data on individual assessments are not readily available. We leave further developments along these lines to future research.

5. Simulation Study

The purpose of this small simulation is to illustrate the parameter recovery of the RSC model using the estimating equations presented in "Appendix." We consider both ML and MAP

estimation and illustrate the benefits of the latter when the individual assessment or the group assessment has few items. Readers interested in replicating our results or conducting further simulations can use the `R` package `scirt` and accompanying documentation, available at www.github.com/peterhalpin/scirt.

The simulation used 2PL IRFs for the items, and the following data-generating parameters:

$$\alpha_i \sim \text{Uniform}(.6, 2.5) \quad \beta_i \sim N(0, 1.3); \quad \theta_j \sim N(0, 1); \quad v_k \sim N(0, 1)$$

The distributions of the item parameters (α_i and β_i) were chosen to realistically reflect the empirical example described in the following section. The standard normal prior on v_k ensures that most weight parameters were in the range $w_k \in [.05, .95]$ (i.e., $v_k \in [-3, 3]$). The total number of items generated was $I = 200$, with half used on the individual assessment and half on the group assessment. To simulate shorter assessments, a single random sample of 20 items was selected from each assessment; this shorter test length also corresponds to our empirical example. Data were simulated for $J = 1000$ respondents, which were matched at random into $K = J/2$ non-overlapping pairs. Parameter recovery and precision were examined in each of four conditions obtained by crossing the long (100 items) and short (20 items) test forms of the individual and group assessments.

Figure 1 summarizes the parameter recovery of the two estimators in each of the four test length conditions. The gray line represents perfect agreement between the estimates and the data-generating values, and the dashed and solid lines represent the loess-smoothed ML and MAP estimates, respectively. Both estimators perform reasonably on the long group test, but when the group test involved a small number of items, bias was apparent for both estimators. Additionally, the bias tended to be larger for the ML estimator, especially at larger magnitudes of v_k.

Figure 2 displays the SEs for the two estimators in each of the four conditions. The SEs were computed by inverting the observed Fisher information at the estimated values, as per the estimation procedure described in "Appendix." For the ML estimator, a total number of 305 cases (15%) had very large standard errors. Most of these cases (81%) occurred with the short group test, although some cases also resulted from perfectly correct or incorrect response patterns on the short individual test. These values are omitted in Fig. 2. Unsurprisingly, the standard errors were smaller when the group test was longer, and for the short group test, inference about the performance of many dyads was highly unreliable. Also as expected, the standard errors of the MAP estimates were substantially lower for larger magnitudes of v_k.

This small simulation study has indicated that the proposed estimation procedures operate as expected when the model is correctly specified. The longer test conditions illustrate parameter recovery in ideal circumstances, although even here ML performed markedly worse than MAP with values of v_k outside the range $[-2, 2]$. While neither estimator performed very well in the short–short condition, performance under these circumstances is important for informing the limitations of the empirical example. In this condition, the MAP estimator remained unbiased over the range of values that we would expect to recover, although the large standard errors indicate that inference about individual dyads was highly unreliable.

6. Empirical Example

We apply the one-parameter RSC model to address the following focal question: Is there evidence of group synergy when real dyads work together? As preliminary analyses, we also investigated (a) the measurement invariance of items calibrated for individual performance, when used in a group setting, and (b) the goodness of fit of the one-parameter RSC model using a parametric bootstrap of the log-likelihood.

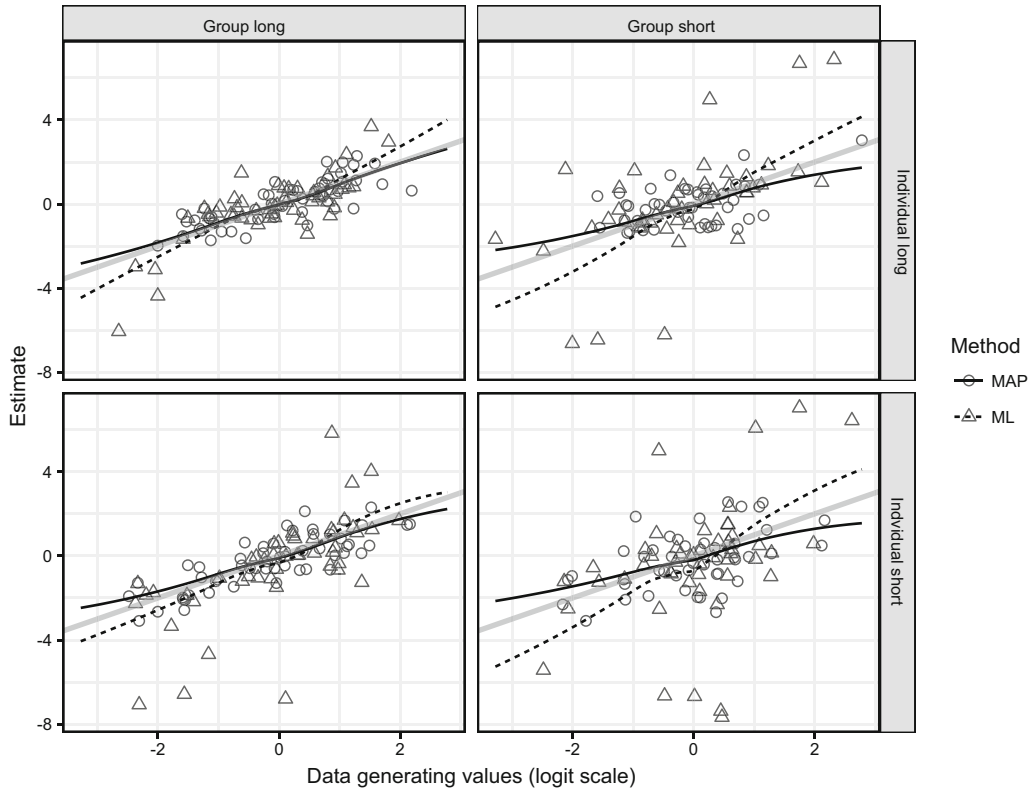

FIGURE 1.
Parameter recovery of v for ML and MAP estimators, in the four test length conditions. The "long" condition denotes the full pool of 100 items, and the "short" condition denotes a random subsample of 20 items, for each of the group and individual assessments. The dashed and solid lines represent the loess-smoothed ML and MAP estimates, respectively. The triangles and circles represent variation in the ML and MAP estimates, respectively, using a random sample of points from each condition.

6.1. Sample and Procedure

Respondents were solicited using Amazon Mechanical Turk (AMT). Approximately 5000 AMT workers were pre-screened using a demographic survey, and these workers constituted the sampling frame for the present study. The sampling frame was comprised of AMT workers who self-reported to live in the USA and to speak English as their first language. The median age was 32 years, with an interquartile range of [27, 40]. The majority of the sampling frame (71%) self-identified as being of "White" ethnicity, 51% reported being female, and 88% reported having at least one year of post-secondary education. Two independent samples were taken from the sampling frame, a calibration sample ($N = 528$) and a research sample ($N = 322$).

The calibration sample was used to estimate item parameters of the 2PL model for a pool of $I = 60$ twelfth-grade mathematics items obtained from previous administrations of the National Assessment of Educational Progress (NAEP). The mathematical content of the items was preserved, but they were modified to be delivered online and to use numeric response rather than multiple-choice formatting. Additionally, participants were instructed to complete the assessment in whatever conditions they deemed suitable, and were explicitly permitted to use a calculator and the Internet. Item parameters of the 2PL model were estimated using maximum likelihood, and a

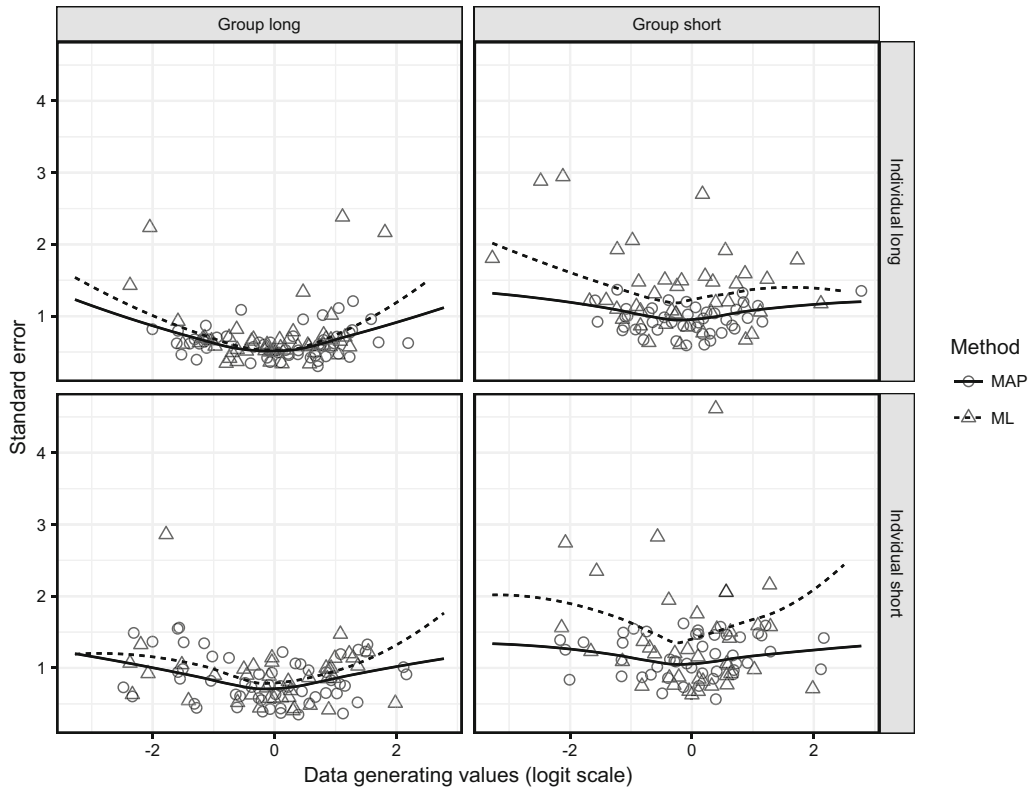

FIGURE 2.
Standard errors (SEs) of the ML and MAP estimates, in the four test length conditions. The "long" condition denotes the full pool of 100 items, and the "short" condition denotes a random subsample of 20 items, for each of the group and individual assessments. The dashed and solid lines represent the loess-smoothed ML and MAP standard errors (SEs), respectively. The triangles and circles represent variation in the ML and MAP SEs, respectively, using a random sample of points from each condition. The SEs were computed using the observed Fisher information.

total of three items were removed from the item pool due to non-monotone IRFs. The remaining items had parameter estimates in the following ranges: $\hat{\beta}_i \in [-3.80, 2.62]$ and $\hat{\alpha}_i \in [0.65, 2.86]$.

In the research sample, all respondents were assessed under both individual and group testing conditions, and the content of the individual and group tests was counterbalanced. In the individual testing condition, respondents were administered a form consisting of 20 items from the calibration sample, which was used to estimate their mathematical ability. After completing the individual assessment, respondents were routed to a second form consisting of another 20 items from the calibration sample. Before commencing the second form, respondents were provided with same instructions as for the individual form, with the exceptions that (a) they would be paired with an anonymous partner, and (b) they were encouraged to work with their partner to ensure that both individuals arrived at the correct response. After acknowledging the instructions, respondents were randomly paired based on their arrival in the routing queue, and they interacted with their partner via online chat.

The online testing platform led to two main limitations in the study design. First, items could not be adaptively administered or even randomized within forms. Second, it was originally intended to match respondents based on their performance on the individual pre-test, thereby making use of the results of Sect. 3. However, it proved to be infeasible to implement anything other than matching based on arrival times. Thus, the order of the individual and group testing

TABLE 2.
Measurement invariance in individual and group testing conditions.

Model	LR	df	p
Metric	36.971	37	.470
Scalar	101.071	74	.020
Scalar (w/ drop)	84.358	70	.116

LR denotes the Satorra–Bentler adjusted likelihood ratio test against the configural model, df its degrees of freedom, and p its right-tail probability. "Scalar (w/ drop)" denotes the scalar invariance model after dropping the two items that were identified by visual inspection of the parameter estimates.

conditions was not counterbalanced, and we were unable to make use of performance on the individual test when selecting partners for the group test. Despite these limitations, we are not aware of any other dataset that provides an opportunity to study small groups working together on calibrated test items.

6.2. Measurement Invariance

Measurement invariance was assessed using the calibration sample and the individual (not conjunctively scored) response patterns of participants in the group testing condition. This resulted in an independent samples design, where respondents in the group testing condition were nested within dyads. Due to the counterbalancing of the individual and group tests, each item in the group testing condition was responded to by only half (161) of the participants in the research sample. Therefore, the following results should be regarded as highly preliminary.

The analysis was implemented in Mplus 7 (Muthén & Muthén, 2017) using the cluster-robust maximum likelihood estimator (Muthén & Satorra, 1995). Respondents in the group testing condition were clustered within dyads, and respondents in the calibration sample each formed their own cluster of one. Measurement invariance was assessed using the Satorra–Bentler adjusted likelihood ratio test (Satorra & Bentler, 2010).

The reader will recall that metric invariance requires equality of factor loadings (discrimination parameters) over groups, and scalar invariance additionally requires equality of item thresholds (difficulty parameters). These models are both nested within the configural model, which places no parameter constraints over groups. As summarized in Table 2, the metric model was retained, but the scalar model was not. Visual inspection of the estimated threshold parameters in the configural and scalar models indicated two items that were plausible sources of invariance. With these two items removed, the scalar model fit reasonably well, as indicated in the last row of Table 2. All subsequent analyses omitted these two items from the group testing condition.

In summary, the measurement invariance analysis indicated that individual and group performance on the mathematics items were largely commensurable. The average mathematics ability was higher in the group testing condition, with a standardized group mean difference of $d = 0.598$ (SE = 0.162). It may be concluded that, on average, respondents performed better when working in dyads than when working individually. However, it is important to note that this analysis does not tell us whether any group exhibited synergy.

6.3. Goodness of Fit

Goodness of fit of the one-parameter RSC model was assessed via parametric bootstrap of the log-likelihood. This amounts to a frequentist version of a posterior predictive model check using Levine and Rubin's (1979) person fit statistic, except applied to the log-likelihood of the RSC model rather than to a more standard IRT model. The RSC model was fitted to the $N =$

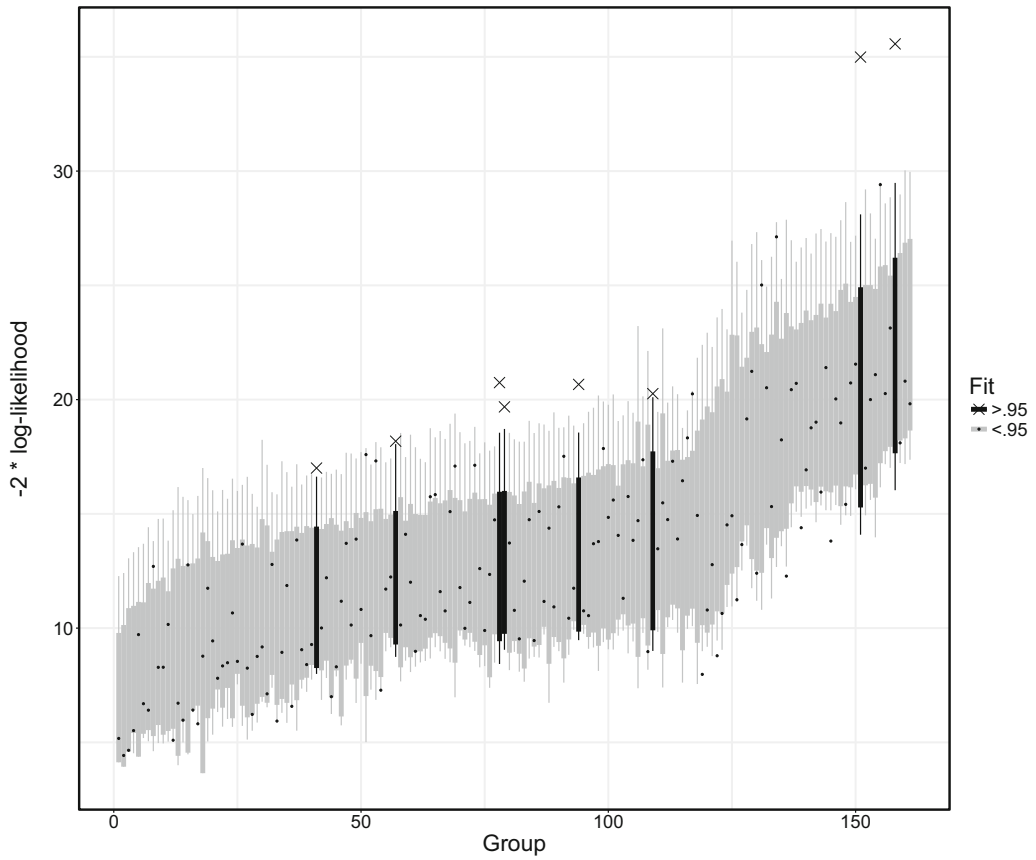

FIGURE 3.
Goodness of fit for each group. Reference distributions for log-likelihood of the one-parameter RSC model were generated using 500 parametric bootstrap replications for each dyad. The figure shows 80% (thick) and 95% (thin) confidence intervals. Groups denoted by black crosses and intervals had fitted log-likelihoods that were improbable under the assumption that the fitted model was the data-generating distribution.

161 conjunctively scored response patterns using the MAP estimator. For each dyad, $R = 500$ replicated responses to the group assessment were generated from the fitted model. The weight of the RSC model was re-estimated for each generated data set, and its log-likelihood was computed (see Eq. (18) in "Appendix"). We then compared the observed value of the log-likelihood to the bootstrapped sampling distribution.

Goodness of fit is summarized in Fig. 3. Eight groups (4.9%) whose fit would be rejected at the 5% (one-tailed) significance level are indicated. These groups were omitted from further analyses. We conclude that the RSC model adequately represented the performance of most groups in the present sample.

6.4. Results

Finally we consider the MAP estimates of v, denoted $\hat{v}$. The results are summarized in Fig. 4. As expected from the data simulation, inference about v was highly unreliable for most groups, due to the short length of both test forms. The marginal reliability (total variance minus the mean of $[\mathrm{SE}(\hat{v})]^2$) was estimated to be .35. Therefore, we simply interpret whether the approximate confidence (credible) interval on $\hat{v}$ included the value of 0.

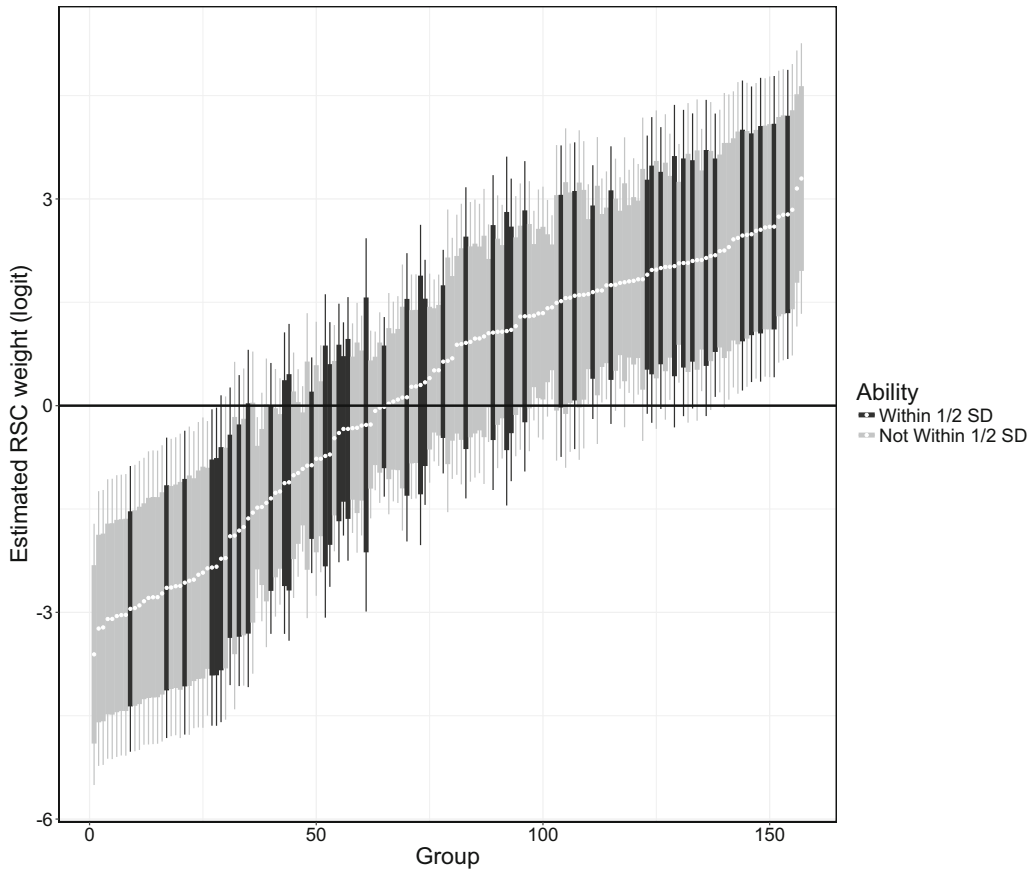

FIGURE 4.
Estimated weights on logistic scale, with approximate 80% (thick) and 95% (thin) confidence intervals for each group. Confidence intervals were computed using a Gaussian reference distribution with the approximate posterior standard deviation of the RSC weight, computed via the observed Fisher information. Groups denoted by black intervals had partners whose ability estimates were within 1/2 standard deviation unit of each other.

For partners with proximate levels of ability, $v > 0$ is evidence of synergy (see Sect. 4.1). In the figure, proximate ability was operationalized as being within 1/2 standard deviation unit on the ability scale. This value was chosen because it is easily interpretable and corresponded to about $2 \times \text{SE}(\hat{\theta})$ for most respondents. From the figure we can see that a total of 6 out of 47 of dyads (13.3 %) whose partner's had proximate ability were also inferred to have exhibited synergy at the 95% confidence level. At the 80% level, this figure increased to 16 (35.5%) of the matched dyads.

In summary, the empirical example has provided initial evidence about the applicability of the one-parameter RSC model to data collected from real dyads. The assumption of measurement invariance of item parameters in individual testing and group testing conditions was shown to be tenable for these data. The model was found to provide acceptable fit for most dyads, although the sensitivity of the model-checking procedure to misspecification is an important topic for future research. As expected from the simulation study, inferences about the decision parameters of individual dyads were highly unreliable given the limitations of the example data. However, the results suggested that some randomly paired dyads who happened to be matched on ability did indeed exhibit evidence of group synergy. Replication of these results under more optimal testing conditions is an important step forward.

7. Conclusions

This paper has shown how the social combination theory of group problem solving can be used to extend existing psychometric models to collaborative settings. We proposed a restricted social combination (RSC) model for pairwise group work under a conjunctive scoring rule for binary (correct / incorrect) item responses. The RSC model was shown to preserve latent monotonicity of the group IRFs and to have a number of relatively intuitive implications for the design of group assessments. In particular, we outlined conditions on team composition and item selection that are necessary for an assessment to provide evidence of group synergy (i.e., better-than-individual performance). Optimal design of group assessments for other purposes and under more general models is an important area of future research.

Because the RSC model is not identified under all team composition conditions, we proposed a one-parameter version of the model as a viable alternative for data analysis and inference. Equations for maximum likelihood and modal a' posteriori estimation of the model were provided (see "Appendix"), and data simulation demonstrated the advantages of the latter with short tests. The real data example provided a preliminary evaluation of a main model assumption (measurement invariance of item parameters in group testing conditions) and the model's fit to real data. We concluded that about 13% of dyads whose members had proximate levels of ability also exhibited evidence of group synergy. Unfortunately, the online testing platform did not allow for the results on assessment design to be put into action. Another priority for future research is the design and implementation of software for delivering group assessments.

Further extensions of this research include models for (a) non-binary group responses, (b) different types of scoring rules, (c) groups with more than two members, and (d) tasks that impose restrictions on group performance (e.g., hidden-profile or jigsaw tasks). Finally, we suggest that models and software that support multiple group memberships will be a major technical challenge to be addressed in order to make inferences about individual-level performance in group contexts. It is intuitive that we should marginalize over groups, rather than conditioning on membership in a single group, when making inferences about group members. Our initial work along these lines suggests that multiple group memberships can be used to identify the two-parameter RSC model and also to identify its parameters at the person level as opposed to the group level. Given ongoing progress in these areas, we hope that group assessments will be a practical reality in the near future.

Acknowledgments

This research was supported by a Spencer Foundation Postdoctoral Fellowship awarded to the first author and a New York University Center for Data Science Seed Grant awarded to both authors.

8. Appendix

8.1. Proofs

This section contains the proofs for Propositions 1 through 7. We let $j = 1, 2$ denote the members of an arbitrary dyad and assume that $\theta_1 \leq \theta_2$ by choice of notation. Subscripts for items are omitted. Several proofs require derivatives of monotonic functions, which the reader will recall are defined almost everywhere on their domain.

8.1.1. Proof of Proposition 1 Let $f, g : \mathbb{R} \to [0, 1]$ be monotone non-decreasing functions, and let $a, b \in [0, 1]$ be fixed constants. The function

$$\begin{aligned} h(x, y) &= a\, f(x)[1 - g(y)] + b\,[1 - f(x)]g(y) + f(x)g(y) \\ &= a\, f(x) + b\, g(y) + (1 - a - b)\, f(x)g(y) \end{aligned} \tag{15}$$

is seen to be non-decreasing in x for fixed y by considering it partial derivative in x and noting that $df/dx = f'(x) \geq 0$:

$$\begin{aligned} \frac{\partial}{\partial x} h(x, y) &= a\, f'(x) + (1 - a - b)\, f'(x)\, g(y) \\ &= a\, f'(x)\,[1 - g(y)] + (1 - b)\, f'(x)\, g(y) \geq 0. \end{aligned} \tag{16}$$

A similar argument shows that Eq. (15) is also non-decreasing in y, and Proposition 1 follows directly.

8.1.2. Proof of Proposition 2 Let $f(x, y) = x(1 - y)$ with $0 < x \leq y < 1$. We show f is strictly concave with global maximum $f(1/2, 1/2) = 1/4$.

A sufficient condition for f to be strictly concave is that $\boldsymbol{u}' H\, \boldsymbol{u} < 0$, where $H = \left(\begin{smallmatrix} 0 & -1 \\ -1 & 0 \end{smallmatrix}\right)$ is the Hessian of f and $\boldsymbol{u} = (u_1, u_2)$ is in the domain of f. The quadratic form reduces to $q = -2\, u_1 u_2$, and the u_i are strictly positive, so $q < 0$.

The global maximum can be found by applying the Karush–Kuhn–Tucker (KKT) conditions for constrained optimization as follows (Boyd & Vandenberghe, 2004, see e.g.). The only inequality that is active at the proposed solution is $g(x, y) = x - y \leq 0$, so the objective function and its gradient may be written, respectively, as

$$\begin{aligned} L(x, y, \mu) &= f(x, y) - \mu\, g(x, y) = x(1 - y) - \mu(x - y), \\ \nabla L(x, y, \mu) &= \begin{bmatrix} 1 - y + \mu \\ -x - \mu \end{bmatrix}. \end{aligned}$$

The KKT conditions state that any local maximum (x^*, y^*) of f must satisfy $\nabla L(x^*, y^*, \mu) = \mathbf{0}$, and $\mu\, g(x^*, y^*) = 0$ for $\mu \neq 0$. These equations are readily solved to show $y^* = x^* = 1/2$.

8.1.3. Proof of Proposition 3 Part 1 of the proposition follows from directly from the definition of θ_0 and the global maximum of $\Delta(P_1, P_2)$ derived in Proposition 2.

Part 2 additionally uses the result (from Proposition 2) that $\Delta(P_1, P_2)$ is strictly concave, and the assumption (from Proposition 3) that $P(\theta)$ is strictly increasing on $\mathcal{N}$, which together imply that $\Delta(u_{12})$ is strictly decreasing in each coordinate of $u_{12} = (\theta_0 - \theta_1, \theta_2 - \theta_0)$, for $\theta_1, \theta_2 \in \mathcal{N}$. The result then follows from writing $\delta = (\theta_2 - \theta_0) + (\theta_0 - \theta_1)$.

8.1.4. Proof of Proposition 4 Let $P(z_j) = [1 + \exp\{-z_j\}]^{-1}$ with $z_j = \alpha(\theta_j - \beta)$ and $Q(z_j) = 1 - P(z_j)$. We show that

$$\arg\max_{\beta} \{P(z_1)\, Q(z_2)\} = (\theta_1 + \theta_2)/2.$$

First note that

$$\frac{\partial}{\partial \beta} P(z_1)\, Q(z_2) = \alpha\, P(z_1)\, Q(z_2)\, [P(z_2) - Q(z_1)].$$

Setting this to zero gives

$$Q(z_1) = P(z_2) \quad \Leftrightarrow \quad P(-z_1) = P(z_2) \quad \Leftrightarrow \quad -z_1 = z_2, \tag{17}$$

hence there is a single critical point at $\beta^* = (\theta_1 + \theta_2)/2$. To show that this is a local maximum, we first find the second derivative,

$$\frac{\partial^2}{\partial \beta} P(z_1)\, Q(z_2) = \alpha^2\, P(z_1)\, Q(z_2) \Big([P(z_2) - Q(z_1)]^2 - P(z_1)\, Q(z_1) - P(z_2)\, Q(z_2) \Big),$$

then use Expression (17) to write

$$\begin{aligned} \frac{\partial^2}{\partial \beta} P(z_1)\, Q(z_2) \Big|_{\beta^*} &= \alpha^2\, P(z_1)^2 \Big([Q(z_1) - Q(z_1)]^2 - 2P(z_1)\, Q(z_1) \Big) \\ &= -2\,\alpha^2\, P(z_1)^3\, Q(z_1) < 0. \end{aligned}$$

Since there is only a single critical point and this is a local maximum, it follows that β^* must also be the global maximum that $P(z_1)\, Q(z_2)$ is strictly concave in β.

8.1.5. Proof of Proposition 5 Using the same notation as above, let $z_j^* = \alpha(\theta_j - \beta^*)$. We show that $P(z_1^*)\, Q(z_2^*)$ is monotone non-increasing in α as follows:

$$\frac{\partial}{\partial \alpha} P(z_1^*)\, Q(z_2^*) = \frac{\partial}{\partial \alpha} [P(z_1^*)]^2 = 2\,(\theta_1 - \beta^*)\, P(z_1^*)\, Q(z_1^*) \leq 0.$$

The first equality uses Expression (17), and the inequality follows since $\theta_1 \leq \beta^*$ by choice of subscripts $j = 1, 2$.

8.1.6. Proof of Proposition 6 Using the same notation as above, the result

$$\alpha = \frac{2}{\delta}\, \ln \frac{1 - \sqrt{D}}{\sqrt{D}}.$$

follows from using the following equalities to solve for α

$$D = \Delta^*(\theta_{12}) = P(z_1^*)\, Q(z_2^*) = [P(z_1^*)]^2.$$

8.1.7. Proof of Proposition 7 Part 1 of the proposition requires computing the Fisher information of a, which is obtained by writing the Bernoulli density of $Y_i \in \{0, 1\}$ as $f(Y_i \mid \zeta) = R_i^{y_i} + (1 - R_i)^{(1-y_i)}$ with R_i defined as in Eq. (9):

$$R_i = P_{i1}\, P_{i2} + a\, P_{i1}\, Q_{i2} + b\, Q_{i1}\, P_{i2}.$$

Part 2 uses the result (from Proposition 3) that $\Delta_i = P_{i1} Q_{i2} = 1/4$ if and only if $\theta_1 = \theta_2$. Then $P_{i1} = P_{i2}$ and $R_i = P_{i1}^2 + (a + b) P_{i1}\, Q_{i1}$, which shows that the value of R_i is not affected by exchanging the values of a and b.

8.2. Estimating Equations

This section provides equations for ML and MAP estimation of the one-parameter RSC model. Referring to Sects. 2.1 and 2.2, let θ_r and $\boldsymbol{X}_r$ denote the latent trait and response pattern, respectively, for respondent r. The group response vector is denoted as $\boldsymbol{Y}$, and $v = \text{logit}(w)$ is the logit of the weight from the one-parameter RSC model in Eq. (14). We let $P_{ir} = P_i(\theta_r)$ denote the IRF for item i on an individual assessment, and $R_j = R_j(\boldsymbol{u})$ denote the group IRF for item j on a group assessment, for $\boldsymbol{u} = (\theta_r, \theta_2, v)$. Estimation using the equations outlined in this section is implemented in the `R` package `scirt` available at www.github.com/peterhalpin/scirt.

Using the local independence assumptions for individual and group assessments, the log-likelihood of interest is

$$\ell(\boldsymbol{u} \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{Y}) = \sum_i \ell(\theta_1 \mid X_{i1}) + \sum_i \ell(\theta_2 \mid X_{i2}) + \sum_j \ell(\boldsymbol{u} \mid Y_j) \tag{18}$$

where

$$\ell(\theta_r \mid X_{r1}) = x_{ir} \ \ln(P_{ir}) + (1 - x_{ir}) \ \ln(1 - P_{ir})$$

and

$$\ell(\boldsymbol{u} \mid Y_j) = y_j \ \ln(R_j) + (1 - y_j) \ \ln(1 - R_j).$$

Methods for estimating θ_r via $\ell(\theta_r \mid X_{ir})$ are well known (Baker & Kim, 2004, e.g.), so we focus on estimation of v via $\ell = \ell(\boldsymbol{u} \mid Y_j)$. Its gradient is

$$\nabla \ell = \frac{\partial}{\partial \boldsymbol{u}} \ell = \sum_j m_j \left[\frac{\partial}{\partial \theta_1} R_j \quad \frac{\partial}{\partial \theta_2} R_j \quad \frac{\partial}{\partial v} R_j \right]^T \tag{19}$$

where

$$m_j = \frac{y_j}{R_j} - \frac{1 - y_j}{1 - R_j}.$$

Letting $P'_{ir} = \frac{\partial}{\partial \theta_r} P_{ir}$ and $w' = \frac{\partial}{\partial v} w$, for $w = \text{logistic}(v)$ the derivatives of the group IRFs in Eq. (9) can be written as

$$\frac{\partial}{\partial \theta_r} R_j = (w + (1 - 2w) \ P_{js}) \ P'_{ir}$$

and

$$\frac{\partial}{\partial v} R_j = (P_{jr} Q_{js} + Q_{jr} P_{js}) \ w'.$$

Let $H(\ell) = \{h_{rs}\}$ denote the Hessian of ℓ, with elements given by

$$h_{rs} = \frac{\partial^2}{\partial u_r \partial u_s} \ell = m_j \ \frac{\partial^2}{\partial u_r \partial u_s} R_j - n_j \ \frac{\partial}{\partial u_r} R_j \ \frac{\partial}{\partial u_s} R_j \qquad r, s = 1, 2, 3 \tag{20}$$

with

$$n_j = \frac{y_j}{R_j^2} + \frac{1 - y_j}{(1 - R_j)^2}.$$

Also let $P''_{ir} = \frac{\partial}{\partial \theta_r} P'_{ir}$ and $w'' = \frac{\partial}{\partial v} w'$. Then the necessary second derivatives are

$$\begin{aligned}
\frac{\partial^2}{\partial \theta_r^2} R_j &= (w + (1 - 2w)\, P_{js})\; P''_{jr} \\
\frac{\partial^2}{\partial \theta_r \partial \theta_s} R_j &= (1 - 2w)\; P'_{jr}\; P'_{js} \\
\frac{\partial^2}{\partial \theta_r \partial v} R_j &= (1 - 2P_{js}) P'_{jr}\; w' \\
\frac{\partial^2}{\partial v^2} R_j &= (P_{jr} Q_{js} + Q_{jr} P_{js})\; w''.
\end{aligned}$$

ML estimation of v can proceed using Eqs. (18) through (20) and the provided derivatives, with standard errors computed by inverting either the observed or expected Hessian. In the latter case, the terms m_j vanish under expectation, and the standard errors can be obtained using only the first-order derivatives of the individual and group IRFs.

In order to demonstrate the identification the weight w, we assume θ_1 and θ_2 are known and compute the Hessian of a single item for v:

$$\frac{\partial^2}{\partial v^2} \ell(\boldsymbol{u} \mid Y_j) = m_j\; (P_{jr} Q_{js} + Q_{jr} P_{js})\; w'' - n_j\; [(P_{jr} Q_{js} + Q_{jr} P_{js})\; w']^2. \tag{21}$$

Setting $v = w$, then $w'' = 0$ and the first term vanishes. The second term is non-positive and equals zero only if $P_{jr} = 0$ or $P_{jr} = 1$ for both $r = 1$ and $r = 2$. Demonstrating identification for $v = \text{logit}(w)$ is less straightforward, but an asymptotic argument shows that $E(m_j) = 0$, in which case the item information again reduces to second term in Eq. (21).

When considering MAP rather than ML estimation, the likelihood in (18) is replaced by the posterior distribution of $\boldsymbol{u}$,

$$p(\boldsymbol{u} \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{Y}) \propto p(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{Y} \mid \boldsymbol{u}) \times p(\boldsymbol{u}). \tag{22}$$

As described in the main paper, we assume that $p(\boldsymbol{u}) = \prod_k p(u_k)$ with $\theta_r \sim N(0, 1)$ and $v \sim N(0, \sigma_v)$. MAP estimation of w proceeds by using

$$\nabla \ell + \frac{\partial}{\partial \boldsymbol{u}} \ln p(\boldsymbol{u}) \quad \text{and} \quad H(\ell) + \frac{\partial^2}{\partial \boldsymbol{u} \partial \boldsymbol{u}^T} \ln p(\boldsymbol{u})$$

in place of Eqs. (19) and (20).

References

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research*, *64*(1), 1–35.

Davis, J. H. (1973). Group decision and social interaction: A theory of social decision schemes. *Psychological Review*, *80*(3), 97–125.

Davis, J. H. (1992). Some compelling intuitions about group consensus decisions, theoretical and empirical research, and interpersonal aggregation phenomena: Selected examples, 1950–1990. *Organizational Behavior and Human Decision Processes*, *52*, 3–38.

Fiore, S. M., Graesser, A., Greiff, S., Griffin, P., Gong, B., Kyllonen, P., et al. (2017). *Collaborative problem solving: Considerations for the National Assessment of Educational Progress*. National Center for Educational Statistics, Washington, DC: Technical report.

Griffin, P., & Care, E. (2015). *Assessment and teaching of 21st century skills: Methods and approach*. New York, NY: Springer.

Heckman, J. J., & Kautz, T. (2014). *Fostering and measuring skills: Interventions that improve character and cognition*. Working Paper No. 19656. Cambridge, MA: National Bureau of Economic Research.

Herman, J., & Hilton, M. (2017). *Supporting students' college success: The role of assessment of intrapersonal and interpersonal competencies.* Washington, DC: The National Academies Press.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, *14*(4), 1523–1543.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272.

Larson, J. R. (2010). *In search of synergy in small group performance*. New York, NY: Taylor & Francis Group.

Laughlin, P. R. (1980). Social combination processes of cooperative, problem-solving groups as verbal intellective tasks. In M. E. Fishbein (Ed.), *Progress in social psychology* (pp. 127–155). Hillsdale, NJ: Erlbaum.

Laughlin, P. R. (2013). *Group problem solving*. Princeton, NJ: Princeton University Press.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, *4*, 269–290.

Lippman, L. H., Ryberg, R., Carney, R., & Moore, K. A. (2015). *Key "soft skills" that foster youth workforce success: Toward a consensus across fields*. Child Trends Publication #2015–24. Washington, DC: Child Trends, Incl.

Lorge, I., & Solomon, H. (1955). Two models of group behavior in the solution of eureka-type problems. *Psychometrika*, *20*(2), 139–148.

Mathieu, J. E., Maynard, T. M., Rapp, T., & Gilson, L. (2008). Team effectiveness 1997–2007: A review of recent advancements and a glimpse into the future. *Journal of Management*, *34*(3), 410–476.

McGrath, J. E. (1984). *Groups: Interaction and performance*. Englewood Cliffs, NJ: Prentice Hall.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 216–316.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus 8 [computer software]*. Los Angeles, CA: Muthén & Muthén.

National Research Council. (2011). *Assessing 21st century skills*. Washington, DC: The National Academies Press.

National Research Council. (2015). *Measuring human capabilities*. Washington, DC: The National Academies Press.

OECD. (2017). *PISA 2015 results, volume V: Collaborative problem solving*. Paris: PISA, OECD Publishing.

Pellegrino, J. W., & Hilton, M. L. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: National Academies Press.

Reckase, M. (2009). *Multidimensional item response theory*. New York: Springer.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled chi square test statistic. *Psychometrika*, *75*(2), 243–248.

Shiflett, S. (1979). Toward a general model of small group productivity. *Psychological Bulletin*, *86*(1), 67–79.

Smoke, W. H., & Zajonc, R. B. (1962). On reliability of group judgements and decisions. In J. H. Criswell, H. Solomon, & P. Suppes (Eds.), *Mathematical Methods in Small Group Processes* (pp. 322–333). Stanford, CA: Stanford University Press.

Steiner, I. D. (1972). *Group processes and productivity*. New York, NY: Academic Press.

von Davier, A., Kyllonen, P., & Zhu, M. (2017). *Innovative assessments of collaboration*. New York, NY: Springer.

Webb, N. M. (1995). Group collaboration in assessment: Multiple objectives, processes, and outcomes. *Educational Evaluation and Policy Analysis*, *17*(2), 239–261.

Manuscript Received: 13 JUN 2017
Final Version Received: 20 JUN 2018
Published Online Date: 9 AUG 2018