# European Journal of Training and Development

Propensity score analysis: an alternative statistical approach for HRD researchers

Greggory L. Keiffer Forrest C. Lane

Received 29 June 2015
Revised 26 February 2016
Accepted 25 April 2016

# Propensity score analysis: an alternative statistical approach for HRD researchers

Greggory L. Keiffer
*University of Texas at Tyler, Tyler, Texas, USA, and*
Forrest C. Lane
*Sam Houston State University, Huntsville, Texas, USA*

## Abstract

**Purpose** – This paper aims to introduce matching in propensity score analysis (PSA) as an alternative statistical approach for researchers looking to make causal inferences using intact groups.

**Design/methodology/approach** – An illustrative example demonstrated the varying results of analysis of variance, analysis of covariance and PSA on a heuristic data set. The three approaches were compared by results and violations of statistical assumptions.

**Findings** – Through the illustrative example, it is demonstrated how different statistical approaches can produce varied results. Only PSA mitigated pre-existing group differences without violating the assumption of independence.

**Originality/value** – This paper attempts to answer calls in the literature for more robust statistical methodologies to better inform human resource development practice and theory.

**Keywords** Human resource development, Theory development, ANCOVA, Intact groups, Propensity score analysis, Statistical assumptions

**Paper type** Research paper

## Introduction

Creation of knowledge in human resource development (HRD) is paramount to the continued advancement of this emerging field. It has been argued that this research must be grounded on the existing literature and must be based on sound methodological practices (Holton, 2002; Reio, 2010a, 2010b). A recent special issue for *Advances in Human Resources Development* described the knowledge gap in HRD concerning the appropriate use of advanced statistical research methods (Reio *et al.*, 2015). These concerns were echoed in Kaufman (2012), who provided an overall dismal rating for the related field of human resource management. Through a perspective article on human resource literature dating back 30 years, it was suggested that "academic research is seriously flawed and inaccurate in its theory" (Kaufman, 2012, p. 14). As a result, there exists a need to help HRD researchers and practitioners to better identify and utilize best practices from the broader academic literature.

This article attempts to answer calls to action for more robust methodology in HRD literature by elucidating an emerging methodological approach available to scholars and practitioners. Propensity score analysis (PSA; Rosenbaum and Rubin, 1983) is an approach that uses information about individuals to make intact groups (e.g. self-selected or

The author(s) received no financial support for the research, authorship and/or publication of this article.

European Journal of Training and Development
Vol. 40 No. 8/9, 2016
pp. 660-675
© Emerald Group Publishing Limited
2046-9012
DOI 10.1108/EJTD-06-2015-0046

non-randomly assigned groups) more directly comparable. Such comparisons may be important for HRD researchers interested in evaluating the effectiveness of an intervention on desired outcomes. Because the aim of this article is to illustrate how PSA may be advantageous over other available approaches for HRD researchers, the article explores through a simulated data set how analysis of variance (ANOVA) and analysis of covariance (ANCOVA) differ from PSA, specifically propensity score matching, and can produce varied results when comparing non-equivalent groups. It is also demonstrated how the data can violate statistical assumptions of ANCOVA and how PSA can help to mitigate this issue. The audience for the article includes HRD practitioners interested in the statistical methods underpinning the field's theory, researchers interested in controlling for selection bias when making group comparisons and a general audience interested in PSA.

## Non-equivalent group comparisons in human resource development

Researchers in HRD often make group comparisons when examining outcomes such as the effectiveness of employee training, subordinate well-being, team development or self-directed learning. Examples of such studies are easily found across the HRD literature (Gaudine and Saks, 2004; Elo *et al.*, 2014; Raes *et al.*, 2015; Durr *et al.*, 1996), and these comparisons often involve participants that are non-randomly assigned to groups. Groups that are not randomly assigned are non-equivalent, and this design is typically considered quasi-experimental (Campbell and Stanley, 1963; Shadish *et al.*, 2002).

Quasi-experimental studies are common in the social sciences, whereas randomized experimental designs tend to be more feasible in other fields of study (e.g. biology, physics and engineering). Randomized experimental design in HRD may be less feasible because of ethical, economic or physical location reasons (Shadish *et al.*, 2006; Stuart, 2010). For example, employees often have contractual obligations that might not allow for assignment to certain conditions. This can result in the use of intact groups that may not be equivalent or appropriate for comparison. Further, not all statistical analyses produce the same result under certain conditions of the data. This has the potential to negatively impact the HRD theory and practitioner outcomes when differences between statistical approaches and their results are not thoroughly understood or taken into consideration.

ANCOVA is an example of an analysis sometimes used in quasi-experimental research designs (Gaudine and Saks, 2004). The purpose of the analysis is to partition out shared variance between the covariate and the outcome and statistically control for this relationship. The use of ANCOVA may be appropriate when groups are randomly assigned, but it may not be well-suited to control for differences because of intact groups. When used with intact groups, ANCOVA has the potential to mislead both researchers and consumers of research about a study's findings (Lord, 1969; Miller and Chapman, 2001). This is in part because of an underlying assumption of independence in the data, defined as the absence of a relationship between group assignment and covariates (Pedhazur, 1997). The covariates used in the ANCOVA model should be unrelated to group assignment. Clear reporting of the data to meet certain assumptions provides the reader with knowledge of the suitability of the statistical approach deployed by the researchers (Gaudine and Saks, 2004). Unfortunately, the ability of the data to meet these statistical assumptions is not always reported in the literature (Elo *et al.*, 2014), and there is no clear guidance provided about how to best manage the data when violations to statistical assumptions exist.

PSA may be a more appropriate statistical approach for intact groups where a relationship between group assignment and a covariate exist (i.e. violation to the assumption of independence). The problem is that this analysis seems to be less utilized in the HRD literature. A key word search within select HRD publications returned only four articles for the phase "propensity score" (Table I). In comparison, searches for the phrases "ANOVA" and "ANCOVA" returned more results. A secondary search for published articles was also conducted through Google Scholar® using the date range of 2005-2015 and search keys "HRD" and "Propensity Score". This broader search also returned scant relevant articles related to PSA with two overlapping the within-journal search results (Lane and Gibbs, 2015; Reio *et al.*, 2015; Choi and Kim, 2012). These findings may demonstrate a need for greater exposure and understanding of PSA and its potential benefits when working with intact groups in the HRD literature.

## Propensity score analysis

PSA (Rosenbaum and Rubin, 1983) is a mathematical approach to causal inference, grounded in the Rubin counterfactual framework (West and Thoemmes, 2010), that uses information about individuals (i.e. covariates) to estimate a participant's likelihood of group assignment (propensity score). A propensity score ($\pi$) is defined in Rosenbaum and Rubin (1983) as "the conditional probability ($P$) of assignment to a particular treatment ($T$) given a vector of observed covariates ($X$)" and is expressed as:

$$\pi_i(X_i) = P(T_i = 1|X_i). \tag{1}$$

The predicted probabilities of group assignment represent a composite score for all covariates and can be used as a covariate adjustment, regression weight or variable for matching. A number of studies have reported differences in results as a function of using PSA (Dehejia and Wahba, 2002; Morgan and Harding, 2006; Schafer and Kang, 2008). When PSA is well implemented, it has been shown that the results obtained through this analysis are more comparable to those obtained through a randomized experimental design (Luellen *et al.*, 2005).

The aim of PSA is to balance a study on the observed covariates and better approximate the expected output from a randomized experimental design (Rubin, 1997). The analysis is summarized briefly here as consisting of four major steps:

1. modeling;
2. conditioning;
3. balancing when a matching approach to PSA is used; and
4. estimation of effects on the resulting matched sample (Lane and Gibbs, 2015).

**Table I.** Summary of the number of articles published in HRD journals using ANOVA, ANCOVA or PSA

| | Date | Total # | Keyword search | | |
|---|---|---|---|---|---|
| Journal searched | range | articles | "ANOVA" | "ANCOVA" | "PSA" |
| *Human Resource Development Quarterly* | 1990-2015 | 961 | 4 | 1 | 0 |
| *Human Resource Development International* | 1998-2015 | 758 | 0 | 0 | 0 |
| *Human Resource Development Review* | 2002-2015 | 339 | 2 | 2 | 1 |
| *Advances in Developing Human Resources* | 1999-2015 | 638 | 10 | 4 | 2 |
| *European Journal of Training and Development* | 1977-2015 | 1,986 | 27 | 4 | 1 |
| Total | | 4,682 | 43 | 11 | 4 |

First, information likely to influence group selection (covariates) should be identified and used in the estimation of propensity scores (modeling). Once an adequate covariate pool has been identified, the probability of group selection (propensity scores) is estimated using these covariates. The use of logistic regression tends to be the most commonly reported method (Guo and Fraser, 2010), but these probabilities may be estimated through other approaches such as probit regression or classification trees (Austin, 2007). Propensity scores are then conditioned as part of the analysis. Matching on propensity scores is the most commonly reported conditioning approach (Thoemmes and Kim, 2011). When matching is used, the aim of the analysis is then to produce groups that share approximately the same probability of group assignment, replicating conditions of random assignment (balancing). Within a region of common support or shared overlap in the distribution of propensity scores between groups, participants are matched, whereas those that do not have similar likelihoods are discarded. The adequacy of the retained matches is examined by assessing the equality of propensity scores between groups pre- and post-matching. The standardized difference in the mean propensity score between groups should be near zero post-matching (Rubin, 2001). Matched groups can then be more directly compared on the outcome of interest.

Because HRD researchers may be likely to use ANOVA-type designs over PSA when making group comparisons, three different approaches to the data are illustrated through the use of a heuristic dataset. The three approaches included a null model where the use of intact groups was ignored (ANOVA), an ANCOVA with one covariate and a PSA model where the same covariate was used to match participants across groups. Through these analyses, it is revealed how statistical outcomes varied with each approach and the impact on the relationship between the covariate and group assignment (i.e. assumption of independence).

It is acknowledged that a PSA model with only one covariate can be unrealistic. A systematic review of educational and psychology research indicated that the median number of covariates used in the estimation of propensity scores was 16 (Thoemmes and Kim, 2011). The concern was that such an example may not be easily interpretable in the context of other analyses (e.g. ANCOVA), particularly among those who want to better understand the underlying logic behind PSA. For example, just three covariates and one grouping (assignment) variable would result in 4 main effects and 11 interaction effects. Thus, the illustrative example has been kept to one covariate to better serve the intended purpose of the paper and to make differences between analyses more accessible to readers. Further, other illustrations of propensity score matching and output with multiple covariates can be found in the literature (Lane and Gibbs, 2015; Lane *et al.*, 2012), although they do not necessarily detail why ANCOVA-type designs may be inappropriate, given underlying relationships within the data. It is hoped that this paper can bridge the gap between those illustrations with a more fundamental understanding of the difference between matching and covariance adjustment.

## Illustrative example

To illustrate the different analytical approaches, an example is presented using a training intervention scenario. Hypothetically, an organization is interested in exploring how an employee's overall compliance adherence was impacted following a compliance training intervention provided to its managers. To examine the research question and illustrate potential approaches to the data, three variables were simulated using the *MASS* package (Venables and Ripley, 2002) in R (v3.2.0):

(1) an interval scale outcome variable of compliance adherence (Y);

(2) an intact treatment group variable (T) where engineering managers ($N$ = 50) received the training and sales managers ($N$ = 50) did not receive the training; and

(3) a single interval scale covariate of employee tenure (X) related to the outcome (i.e. a covariate as defined in this illustration).

Employee tenure was selected as a covariate because we suggest that most companies are likely to have this information prior to the design of an intervention. For simplicity and ease of interpretation, continuous variables were created as $T$-scores with a mean of 50 and a standard deviation of 10. Data were specified to be multivariate normal with an empirical distribution. Variable means, standard deviations and correlations for this data are reported in Table II. The syntax to replicate all data is also available in the Appendix. All analyses were conducted in R (v3.2.0), and both tests of statistical significance and model effect sizes were considered in the interpretation of statistical model results (Wilkinson and APA Task Force on Statistical Inference, 1999).

### *Analysis of variance (null) model*

A null or naïve model was tested first by comparing the engineering manager group to the sales manager group through a one-way ANOVA. This model illustrated a possible result when pre-existing differences, due to the use of intact groups, is ignored. The naïve model may be unlikely in practice but may help to better differentiate ANOVA from other statistical approaches.

The results of the null model ANOVA are reported in Table III. Prior to interpreting those results, the assumption of homogeneity of group variances was examined among the data. Levene's test indicated that group variances were similar ($F$[1,98] = 0.416, $p$ = 0.520). Given this result, the ANOVA model with one grouping variable was interpreted further, and findings suggested a higher level of compliance adherence for the

**Table II.**
Correlation matrix of variables among all data in the unmatched and matched samples

| | Unmatched ($N$ = 100) | | | Matched ($N$ = 32) | | |
|---|---|---|---|---|---|---|
| Variable | Y | T | X | Y | T | X |
| Y | – | | | – | | |
| T | 0.382 | – | | 0.395 | – | |
| X | 0.436 | 0.729 | – | 0.287 | 0.088 | – |
| *M* | 50.000 | 0.460 | 50.000 | 50.500 | 0.500 | 50.160 |
| *SD* | 10.000 | 0.500 | 10.000 | 8.470 | 0.510 | 3.820 |

**Table III.**
ANOVA model summary

| Source | df | SS | MS | $F$ | $p$ | $\eta^2$ |
|---|---|---|---|---|---|---|
| Effect | 1 | 1,447 | 1,446.6 | 16.77 | <0.001 | 0.146 |
| Residual | 98 | 8,453 | 86.3 | | | |
| Total | 100 | 9,900 | | | | |

engineering manager group ($M$ = 54.12, $SD$ = 8.18) compared to the sales manager group ($M$ = 46.49, $SD$ = 10.13). This mean difference was statistically significant ($F$[1,98] = 16.77, $p < 0.001$), and group assignment explained approximately 15 per cent of individual differences in the outcome (compliance adherence) based on the eta squared ($\eta^2$) effect size. The specific criteria for determining a meaningful result may vary based on a number of factors (e.g. specific theory tested, prior research findings, etc.). For the purpose of illustration, the result was interpreted to be meaningful, given the group means were statistically different from one another and resulted in a modest effect size. Some researchers might conclude that the compliance training intervention utilized by the organization resulted in an increased level of compliance adherence.

The result from the ANOVA may be encouraging, but it can also be prone to bias if pre-existing differences were not considered. Such differences may stem from the use of intact groups such as managers from disparate divisions. Depending on the nature and content of the training, some managers may have been more or less likely to have received the training and to adhere to compliance policies. For example, a training intervention on budgeting policies may have had more appeal to managers in business management (e.g. sales) than those in more technical operations (e.g. engineering), which can confound the result. Unfortunately, the magnitude of this bias and its effect on the interpretation of the statistical result is unlikely to be known unless these differences are considered in the design of the study.

### *Analysis of covariance model with one covariate*

Relevant covariates may be included in the design of a study to mitigate threats to its validity (Shadish *et al.*, 2002). A covariate is defined generally as a source of experimental error related to the outcome (Hinkle *et al.*, 2003). Liao *et al.* (2014) suggested that covariates may "include country variables such as cultural values and GDP, organizational variables such as technology level and organizational size, and individual variables such as personality and education level" (p. 130). By partitioning out the relationship between these variables and the outcome (e.g. compliance adherence), the use of ANCOVA may result in a smaller error variance and a more powerful test of statistical significance when comparing groups (Hinkle *et al.*, 2003).

To illustrate the use of ANCOVA in the data, employee tenure (X) was included as a covariate and was added to the initial ANOVA model. The covariate was entered into the model first to "statistically control" for this relationship prior to interpreting the treatment of training effect. The result of this analysis suggested the covariate, employee tenure (X) of managers, was statistically related to compliance adherence ($F$[1,96] = 23.787, $p < 0.001$) and shared 19 per cent of variance in common (Table IV). This finding supported the use of this variable as a covariate in the analysis. Further, the residual or sum of squared error variance in the ANCOVA model was reduced from

| Source | df | SS | MS | $F$ | $p$ | $\eta^2$ |
|---|---|---|---|---|---|---|
| Employee tenure (X) | 1 | 1,882 | 1,882.0 | 23.787 | <0.001 | 0.190 |
| Engineer/Sales Manager (T) | 1 | 88 | 88.1 | 1.066 | 0.304 | 0.008 |
| Interaction (X × T) | 1 | 1 | 1.4 | 0.017 | 0.897 | <0.001 |
| Residual | 96 | 7,930 | 82.6 | | | |
| Total | 99 | 9,900 | | | | |

**Table IV.** ANCOVA model summary

8,453 in the ANOVA model to 7,930 in the ANCOVA model and seemed to indicate the desired result from this statistical approach.

Before interpreting the main effect of group assignment (T) in the ANCOVA model, the data were also examined for statistical differences in the relationship between employee tenure (X) and compliance adherence (Y) based on group assignment (i.e. homogeneity of regression slopes). An assumption in ANCOVA is that one common slope best fits the data. If more than one slope fits the data, there is a violation to the assumption of homogeneity of regression slopes, and the single slope may not adequately explain differences between groups. The homogeneity of regression slopes assumption was tested by including an interaction effect (X*T) in the ANCOVA model. The effect was not statistically significant ($F[1,96] = 0.017, p = 0.897$) and suggested no violation to this assumption.

The common regression slope was estimated to be $b = 0.354$, and this slope (weight) was used to compute an adjusted mean score (i.e. mean adjusted for differences in the covariate), as is typical for this analysis. Adjusted mean scores were defined as:

$$\bar{Y}_{j(adj)} = \bar{Y}_j - b\left(\bar{X}_j - \bar{X}\right) \quad (2)$$

where $\bar{Y}_{j(adj)}$ is the adjusted mean of treatment j; $\bar{Y}_j$ is the mean of treatment $j$ before the adjustment; $b$ is the common regression coefficient; $\bar{X}_j$ is the mean of the covariate for treatment $j$; and $\bar{X}$ is the grand mean of the covariate. The result was an adjusted treatment group mean of 48.74 and an adjusted control group mean of 51.48. By statistically partitioning out the common variance between the covariate and outcome, the authors were able to determine the residual relationship or the supposed true effect due to manager training. Through the use of ANCOVA, engineering managers and sales managers were determined to be the same in their level of compliance adherence ($F[1,88] = 1.08, p = 0.30, \eta^2 < 0.001$), and this resulted in a different interpretation over the null ANOVA model (Table V).

The use of ANCOVA may appear to be a more robust approach for making group comparisons, but it can be problematic when a covariate remains correlated to group assignment. Evidence of such a relationship is a violation to a second assumption of ANCOVA, the independence between the covariate and group assignment. The use of ANCOVA adjusted for the relationships between the covariate and the dependent variable but did not account for the relationship between the covariate and group assignment. When the covariate and group assignment are not statistically independent (i.e. correlated), ANCOVA "would in fact remove the treatment sum of squares part of

**Table V.** Summary of the mean differences and tests of statistical significance by model

| Model | M (Treated) | SD (Treated) | M (Control) | SD (Control) | Difference | $t$ | $p$ | $d$ |
|---|---|---|---|---|---|---|---|---|
| Unmatched ANOVA | 46.48 | 10.13 | 54.12 | 9.18 | 7.64 | 4.17 | <0.01 | 0.82 |
| ANCOVA | 48.74[a] | 1.52[b] | 51.48[a] | 1.69[b] | 2.74 | 1.04 | 0.30 | 0.04 |
| Matched ANOVA (PSA) | 47.21 | 8.90 | 53.79 | 6.79 | 6.58 | 2.35 | 0.03 | 0.17 |

**Notes:** [a] Scores reflect adjusted means; [b] scores represent the standard error of the adjusted mean estimate the treatment you really want included" and confound what remains (Maxwell and Delaney, 1990, pp. 382-383). Notice in this heuristic data that the covariate employee tenure resulted in the removal of nearly all of the variance explained in compliance adherence.

A central concern and key argument of this article is the assumption that independence between the covariate and group assignment may be less understood. To help make that difference more clear for readers, homogeneity of regression slopes is a test of a common slope, whereas the assumption of independence is a test of the relationship between the covariate and group assignment from zero. In the example above, the covariate of employee tenure (X) was not only related to group assignment (T) but also shared half of variance in common with it ($r = 0.729$). This shared variance reflects a problem with the design of the study. Participants were not randomly assigned to groups and were more likely to be in one group over the other. Subsequent comparisons between managers remained confounded in that relationship, despite the use of ANCOVA.

The lack of understanding about the assumption of independence between the covariate and group assignment in ANCOVA may suggest a need to distinguish ANCOVA as a method for adjusting group means due to a covariate and as a method to control for pre-existing group differences. ANCOVA can be appropriate under conditions of random assignment where groups are homogenous and covariates reflect information that cannot be controlled by the researcher (e.g. change in health after group assignment). The use of ANCOVA may not be appropriate when the aim is to statistically control for variables that are correlated to non-random assignment. There is no statistical analysis that can fully ascertain how an individual would have performed if the pre-existing differences (e.g. personality, gender, race and ethnicity) were not present. The reality is variables remain correlated with one another and cannot be disentangled through the use of ANCOVA. This point was well argued by Lord (1969) and has continued to be emphasized on by others over the years (Miller and Chapman, 2001; Pedhazur, 1997).

### *Propensity score matching model*

Although no statistical analysis can fully control for pre-existing differences, PSA can help to mitigate relationships between covariates and group assignment. Rather than partition out shared variance due to the covariates, PSA uses the same covariate information to estimate an individual's likelihood of being assigned a particular group (i.e. propensity score), and when matching is used as a conditioning method, it matches that individual to someone in the other group with a similar disposition or probability of being assigned to it. It is important to note that not all individuals will be retained when matching is used in a PSA. Some participants will be considered too different based on the collective combination of covariate information, and only those individuals who can be well matched are retained for group comparison.

To illustrate how the methodology between these two approaches varies, the same data from the ANCOVA model were used in a PSA. First, the covariate of employee tenure (X) was included in a logistic regression to predict group assignment (T), and this resulted in propensity scores ($\pi$) for each individual. The mean propensity score for the engineering manager group (treated) based on their employee tenure was found to be $\pi = 0.77$ compared to a mean propensity score of $\pi = 0.28$ for the sales manager group (untreated). This seemed to indicate considerable differences in the likelihood of being a manager in one division over the other (Figure 1). Under conditions of random assignment, a researcher would expect the propensity score distributions for both groups to perfectly overlap. The distribution of propensity scores for the engineering managers and sales managers in the sample data was not the same.

The means and standard deviations of propensity scores and the covariate scores were also compared and are shown in Table VI. It was suggested in Rubin (2001) that groups are considered appropriate for comparison when the standardized mean differences are small ($d < 0.20$) and non-significant. The group means in this data were statistically different on both the propensity scores ($t[87.56] = 11.17$, $p < 0.01$) and covariate scores ($t[98] = 10.43$, $p < 0.01$). Further, the standardized mean differences were almost ten times the recommended threshold to be considered adequately matched.

To reduce the disparity between groups on propensity scores and the covariate, the *MatchIt* package (Ho *et al.*, 2007) was used to identify participants who were more directly comparable. Matching on propensity scores is automated in the R syntax for

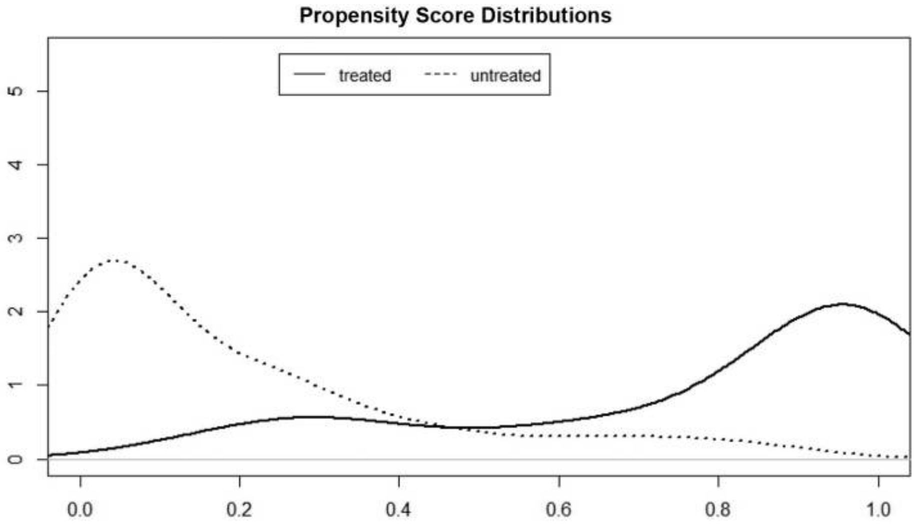

**Figure 1.**
Propensity score distributions prior to matching by group

**Table VI.**
Means, standard deviations and tests of statistical significance pre- and post-matching

| Variable | M (Treated) | SD (Treatment) | M (Control) | SD (Control) | $t$ | $p$ | $d$ |
|---|---|---|---|---|---|---|---|
| *Pre-matching* | | | | | | | |
| Propensity score | 0.77 | 0.28 | 0.19 | 0.23 | 11.17 | <0.01 | 2.07 |
| Covariate | 57.85 | 7.32 | 43.31 | 6.49 | 10.43 | <0.01 | 1.99 |
| $n$ | 54 | 46 | | | | | |
| *Post-matching (caliper = 0.20)* | | | | | | | |
| Propensity score | 0.48 | 0.26 | 0.45 | 0.25 | 0.36 | 0.72 | 0.12 |
| Covariate | 50.49 | 4.01 | 49.83 | 3.71 | 0.48 | 0.63 | 0.09 |
| $n$ (retained) | 16 | | 16 | | | | |
| $n$ (unmatched) | 30 | | 38 | | | | |

users, and details of this process are documented in Ho *et al.* (2007). Propensity scores in the treatment group were sorted and sequentially matched to a single case with the nearest propensity score in the comparison group (i.e. one-to-one matching). The maximum allowable difference for matching must be specified *a priori* in the syntax, and this was done using a caliper of 0.20 standard deviations of the propensity score in the sample data. It has been suggested that a caliper of $d < 0.25$ is reasonable for reducing bias between groups (Stuart, 2010). Cases are retained only if the distance between matched pairs is less than the specified *a priori* caliper distance. More detailed guidance about distance calipers can be found in the literature (Caliendo and Kopeinig, 2008; Thoemmes and Kim, 2011).

The matching algorithm resulted in 16 matches ($n = 32$ cases) that met the criteria for inclusion ($d < 0.20$). The remaining unmatched cases were excluded from further analysis. Matched cases were extracted and compared by group on propensity scores and covariate scores to determine if balance was improved (automated through the *MatchIt* package). The result was a notable improvement in the comparison of pre-existing group differences. In this heuristic example, the standardized mean difference in propensity scores was reduced from an initial group separation of $d = 2.07$ to a post-matching group separation of $d = 0.12$ (Table VI). Similar reductions were found in the covariate scores (employee tenure) between groups where the standardized mean difference was $d = 0.09$ post-matching.

The homogeneity of variance in propensity scores was also compared across groups post-matching. Rubin (2001) suggests that the ratio of the variances in propensity score for both groups should be near one. In the matched data, the ratio of propensity score variances was 1.04, and the ratio of covariate score variances was 1.08. These ratios were an improvement from the results obtained prior to matching and were within recommended limits of the analysis (0.80-1.20) reported in the literature (Rubin, 2001). As a result, both mean likelihoods for group assignment and their distributions were determined to be more directly comparable (Figure 2).

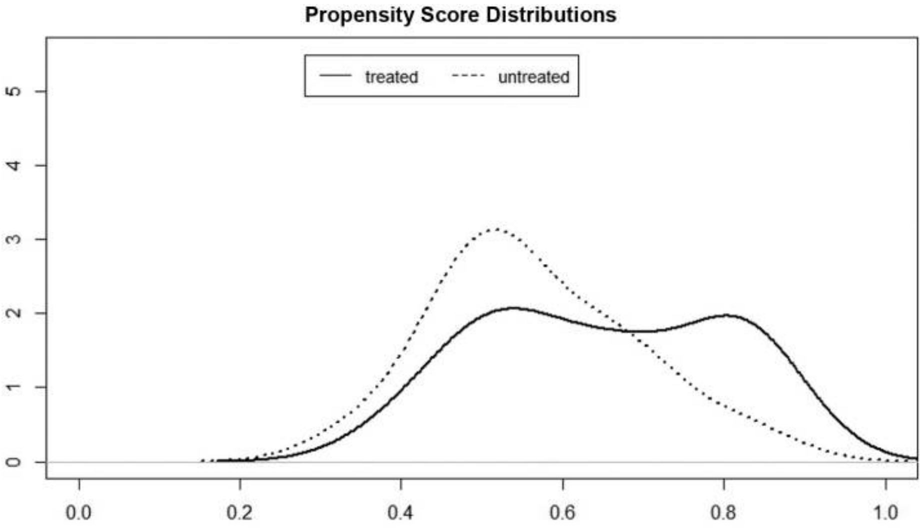

**Figure 2.**
Propensity score distributions after matching by group

It is important to note that PSA mitigated both the pre-existing group differences and the relationship between group assignment (T) and employee tenure (X), the reason it is posited that some may be inclined to use ANCOVA in the first place. Prior to matching, this relationship was high ($r = 0.729$) and resulted in a violation to the assumption of independence in ANCOVA. In the matched subsample, the relationship between group assignment and employee tenure was reduced to near zero ($r = 0.088$). The relationship between these two variables changed because only individuals with similar likelihoods of group assignment were retained post-matching. Pre-existing group differences were mitigated in PSA by removing poorly matched participants. This was in contrast to ANCOVA that retained all individuals and partitioned out only the shared variance between the outcome and covariate, not the variance between the covariate and group assignment.

### *Comparing group differences on matched data*

Having identified a sample of participants where group assignment was unrelated to the covariate, the outcome differences were reexamined by group on the newly matched sample. An advantage to PSA is that once groups are equated on propensity scores and covariates, there is no need to include the covariate(s) in the final group comparison, unless the variable is a true covariate in that it remains correlated to the outcome but unrelated to group assignment and adjusting for differences due to the covariate(s) is the aim of the researcher. The authors compared the matched participants on the outcome of compliance adherence (Y) and determined that the treatment group scored statistically lower on the outcome ($t[30] = 2.35, p = 0.03$). The magnitude of this difference was not as large as suggested by the initial ANOVA for participants (Table V). In this example, the ANOVA that ignored pre-existing group differences (null model) would have overestimated the effect, whereas ANCOVA would have resulted in a type II error, although the specific results obtained here are likely to vary across other studies. Only the matched sample obtained through PSA used a known covariate in a way that did not violate the assumption of independence.

## Discussion and implications

Randomized experimental design may be impractical in the current fluid organizations, and researchers must be keenly aware of potential pitfalls associated with the use of intact groups. If these pitfalls are not well understood, they may lead to results that can violate statistical assumptions and potentially bias results used in HRD theory development. As practitioners rely on academic theory, this places the burden on researchers to use statistical techniques best suited for the design of the study and underlying variable relationships.

A fundamental point argued in this paper is that different statistical approaches can yield varied results on the same data. It is important for researches to understand the sources of this variability. Researchers are likely to be aware that intact groups can be systematically different and, perhaps, it is why some may choose to use ANCOVA or similar designs (e.g. hierarchical regression). However, ANCOVA was not intended to fully mitigate issues associated with non-random assignment. There is no statistical analysis that can fully account for poor study design (Lord, 1969; Miller and Chapman, 2001), and the use of ANCOVA will not necessarily produce the intended result under these conditions. Only through PSA, a subsample that mitigated the relationship between the covariate and group assignment, a problem associated with quasi-experimental design and the use of intact groups, was identified.

Although PSA may provide the HRD academic researcher with a robust statistical alternative for quasi-experimental designs, readers must also be aware of the statistical assumptions and limitations of this analysis. A key assumption of PSA is strongly ignorable treatment assignment. It suggests that propensity scores have been well specified, and results are not adversely affected by unobserved covariates or hidden bias (Rosenbaum, 2010). Only one covariate was illustrated in our heuristic data set for simplicity, but it is highly unlikely that researchers would utilize a single covariate in real-world studies. Best practice would suggest the researcher include multiple covariates in the estimation of propensity scores, as well as pre-test measures of the outcome (Cook and Steiner, 2010).

HRD researchers may have an advantage in the use of PSA given the potential accessibility of covariate data. In fields such as medicine or education, significant challenges may exist in obtaining covariate information. In contrast, HRD researchers and practitioners are likely to work with organizations that employ the individuals they seek to evaluate. Organizations often compile a significant depth of information about their employees that could be used in PSA. This potential availability of employee data may speak to the potential utility of this analysis within HRD research.

PSA models are likely to require sufficiently large samples (Stuart, 2010). Because the method may result in discarded cases that cannot be well matched, the analysis has the potential to impact statistical power. Once again, HRD researchers and practitioners may be at an advantage, given they are often working with a known quantity (i.e. employees at a company). To account for the subsequent impact to statistical power, researchers may want to include more cases prior to matching.

PSA is a multi-faceted analysis and requires good researcher judgment. Use of the analysis does not guarantee that the initially identified subsample of matched data will result in the initial desired reduction to group differences or necessarily balance groups on all covariates used in the analysis. It may be necessary to adjust distance calipers or include additional covariates. Further, there are multiple matching algorithms that may be used to identify adequate matches. Some matching algorithms may be more effective under certain conditions. The authors suggest that interested readers may consult the work of Caliendo and Kopeinig (2008) or Guo and Fraser (2010) for more detailed guidance on PSA.

## Conclusion

The goals of this article were to introduce HRD readers to PSA and compare the analysis to alternative statistical approaches often used when comparing intact groups. Simulated data were used to heuristically guide readers through PSA while making comparisons to ANOVA and ANCOVA on the same data. Through this illustration, the authors demonstrated that the use of ANCOVA with intact groups resulted in violations to the assumption of independence between the covariate and group assignment. Conversely, PSA utilized the same covariate information and resulted in a matched subsample that better met the statistical assumptions of the data. The illustration reinforced that different analyses may produce different results when using the same data containing intact groups. As well-designed studies aim to mitigate threats to validity, HRD researchers who plan to use intact groups may be better served by using PSA.

## References

Austin, P.C. (2007), "The performance of different propensity score methods for estimating marginal odds ratios", *Statistics in Medicine*, Vol. 26 No. 16, pp. 3078-3094.

Caliendo, M. and Kopeinig, S. (2008), "Some practical guidance for the implementation of propensity score matching", *Journal of Economic Surveys*, Vol. 22 No. 1, pp. 31-72.

Campbell, D.T. and Stanley, J.C. (1963), "Experimental and quasi-experimental designs for research on teaching", in Gage, N.L. (Ed.), *Handbook of Research on Teaching*, Rand McNally, Chicago, IL, pp. 171-246.

Choi, H. and Kim, J. (2012), "Effects of public job training programmes on the employment outcome of displaced workers: results of a matching analysis, a fixed effects model and an instrumental variable approach using korean data", *Pacific Economic Review*, Vol. 17 No. 4, pp. 559-581.

Cook, T.D. and Steiner, P.M. (2010), "Case matching and the reduction of selection bias in quasi-experiments: the relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis", *Psychological Methods*, Vol. 15 No. 1, pp. 56-68.

Dehejia, R.H. and Wahba, S. (2002), "Propensity score-matching methods for nonexperimental causal studies", *Review of Economics and Statistics*, Vol. 84 No. 1, pp. 151-161.

Durr, R., Guglielmino, L.M. and Guglielmino, P.J. (1996), "Self-directed learning readiness and occupational categories", *Human Resource Development Quarterly*, Vol. 7 No. 4, pp. 349-358.

Elo, A., Ervasti, J., Kuosma, E. and Mattila-Holappa, P. (2014), "Effect of a leadership intervention on subordinate well-being", *Journal of Management Development*, Vol. 33 No. 3, pp. 182-195.

Gaudine, A.P. and Saks, A.M. (2004), "A longitudinal quasi-experiment on the effects of posttraining transfer interventions", *Human Resource Development Quarterly*, Vol. 15 No. 1, pp. 57-76.

Guo, S. and Fraser, M.W. (2010), *Propensity Score Analysis: Statistical Methods and Applications*, Sage Publications, Thousand Oaks, CA.

Hinkle, D.E., Wiersma, W. and Jurs, S.G. (2003), *Applied Statistics for the Behavioral Sciences*, Houghton Mifflin, Boston, MA.

Ho, D.E., Imai, K., King, G. and Stuart, E.A. (2007), "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference", *Political Analysis*, Vol. 15 No. 3, pp. 199-236.

Holton, E.F. (2002), "The mandate for theory in human resource development", *Human Resource Development Review*, Vol. 1 No. 1, pp. 3-8.

Kaufman, B.E. (2012), "Strategic human resource management research in the United States: a failing grade after 30 years?", *Academy of Management Perspectives*, Vol. 26 No. 2, pp. 12-36.

Lane, F.C. and Gibbs, S. (2015), "Propensity score analysis: a secondary data analysis of work – life policy and performance outcomes", *Advances in Developing Human Resources*, Vol. 17 No. 1, pp. 102-116.

Lane, F.C., To, Y.M., Henson, R.K. and Shelley, K. (2012), "An illustrative example of propensity score matching within education research", *Career and Technical Education Research*, Vol. 37 No. 1, pp. 187-212.

Liao, Y., Sun, J. and Thomas, D.C. (2014), "Cross-cultural research", in Sanders, K., Cogin, J. and Bainbridge, H.T.J. (Eds), *Research Methods for Human Resource Management*, Routledge, New York, NY, pp. 115-135.

Lord, F.M. (1969), “Statistical adjustments when comparing preexisting groups”, *Psychological Bulletin*, Vol. 72 No. 5, pp. 336-337.

Luellen, J.K., Shadish, W.R. and Clark, M.H. (2005), “Propensity scores an introduction and experimental test”, *Evaluation Review*, Vol. 29 No. 6, pp. 530-558.

Maxwell, S.E. and Delaney, H.D. (1990), *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, Wadsworth, Belmont, CA.

Miller, G.A. and Chapman, J.P. (2001), “Misunderstanding analysis of covariance”, *Journal of Abnormal Psychology*, Vol. 110 No. 1, pp. 40-48.

Morgan, S.L. and Harding, D.J. (2006), “Matching estimators of causal effects prospects and pitfalls in theory and practice”, *Sociological Methods & Research*, Vol. 35 No. 1, pp. 3-60.

Pedhazur, E.J. (1997), *Multiple Regression in Behavioral Research: Explanation and Prediction*, Harcourt Brace, Fort Worth, TX.

Raes, E., Kyndt, E., Decuyper, S., Van den Bossche, P. and Dochy, F. (2015), “An exploratory study of group development and team learning”, *Human Resource Development Quarterly*, Vol. 26 No. 1, pp. 5-30.

Reio, T.G. (2010a), “The ongoing quest for theory-building research methods articles”, *Human Resource Development Review*, Vol. 9 No. 3, pp. 223-225.

Reio, T.G. (2010b), “The threat of common method variance bias to theory building”, *Human Resource Development Review*, Vol. 9 No. 4, pp. 405-411.

Reio, T.G., Nimon, K. and Shuck, B. (2015), “Preface: quantitative data-analytic techniques to advance hrd theory and practice”, *Advances in Developing Human Resources*, Vol. 17 No. 1, p. 3.

Rosenbaum, P.R. (2010), “Dilemmas and craftsmanship”, *Design of Observational Studies*, Springer, New York, NY, pp. 3-20.

Rosenbaum, P.R. and Rubin, D.B. (1983), “The central role of the propensity score in observational studies for causal effects”, *Biometrika*, Vol. 70 No. 1, pp. 41-55.

Rubin, D.B. (1997), “Estimating causal effects from large data sets using propensity scores”, *Annals of Internal Medicine*, Vol. 127 No. 8, pp. 757-763.

Rubin, D.B. (2001), “Using propensity scores to help design observational studies: application to the tobacco litigation”, *Health Services and Outcomes Research Methodology*, Vol. 2 Nos 3/4, pp. 169-188.

Schafer, J.L. and Kang, J. (2008), “Average causal effects from nonrandomized studies: a practical guide and simulated example”, *Psychological Methods*, Vol. 13 No. 4, p. 279.

Shadish, W.R., Cook, T.D. and Campbell, D.T. (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton-Mifflin, Boston, MA.

Shadish, W.R., Luellen, J.K. and Clark, M. (2006), “Propensity scores and quasi-experiments: a testimony to the practical side of lee sechrest”, in Bootzin, R.R. and McKnight, P.E. (Eds), *Strengthening Research Methodology: Psychological Measurement and Evaluation*, American Psychological Association, Washington, DC, pp. 143-157.

Stuart, E.A. (2010), “Matching methods for causal inference: a review and a look forward”, *Statistical Science*, Vol. 25 No. 1, pp. 1-21.

Thoemmes, F.J. and Kim, E.S. (2011), “A systematic review of propensity score methods in the social sciences”, *Multivariate Behavioral Research*, Vol. 46 No. 1, pp. 90-118.

Venables, W.N. and Ripley, B.D. (2002), *Modern Applied Statistics with S*, Springer, New York, NY.

West, S.G. and Thoemmes, F. (2010), “Campbell’s and rubin’s perspectives on causal inference”, *Psychological Methods*, Vol. 15 No. 1, pp. 18-37.

Wilkinson, L. and American Psychological Association Task Force of Statistical Inference (1999), “Statistical methods in psychology journals: guidelines and explanations”, *American Psychologist*, Vol. 54 No. 8, pp. 594-604.

## Appendix

```
#Install and load packages to r
install.packages(“effects”)
install.packages(“multcomp”)
install.packages(“MASS”)
install.packages(“MatchIt”)
install.packages(“nonrandom”)
install.packages(“effsize”)
install.packages(“psych”)

#####################
#simulation of data#
#####################
a<−matrix(c(1, .462, .436, .462, 1, .829, .436, .829, 1),3)
sd1<−10
sd2<−.25
sd3<−10
stdevs <− c(sd1, sd2, sd3)
B<−stdevs%*%t(stdevs)
cov<−b*a
library(“MASS”)
set.seed(10001)
simdata<−data.frame(mvrnorm(n = 100, mu=c(50, .5, 50), cov, empirical=TRUE))
names(simdata) = c(“response”,“treatment”,“covariate”)
simdata$treatment[simdata$treatment< .5] <− 0
simdata$treatment[simdata$treatment>=.5] <− 1
simdata$treatment<−factor(simdata$treatment)
library(“psych”)
describe(simdata)
by(simdata, simdata$treatment, describe)

#############################
#Null Model (One-Way ANOVA)#
#############################
#Levene’s test of equal group variances
leveneTest(response ~ treatment, data=simdata, center=“mean”)
ANOVA<−aov(response~treatment, data = simdata)
summary(ANOVA, type = “III”)
library(“effsize”)
cohen.d(response~treatment, simdata)

###################################
#ANCOVA model with one covariate#
###################################
ancova<−lm(response~covariate + treatment + covariate:treatment, data=simdata)
summary(ancova, type=“III”)
library(effects)
ancova<−aov(response~covariate + treatment, data=simdata)
AdjustedMeans<−effect(“treatment”,ancova, se=TRUE)
library(multcomp)
posthocs <− glht(ancova, linfct = mcp(treatment = “Tukey”))
summary(AdjustedMeans)
summary(posthocs)
AdjustedMeans$se

#Test of Independence
check<−aov(covariate~treatment, data = simdata)
summary(check)

#Graphically view propensity score distributions.
library(“nonrandom”)
ps <− pscore(data=simdata, treatment ~ covariate,
       name.pscore=“ps”)
plot.pscore(ps, main=“Propensity Score Distributions”, with.legend=TRUE,
       par.1=list(lty=1,lwd=2), par.0=list(lty=3,lwd=2),
       ylab =“”,ylim=c(0,5.5), xlim=c(0,1.0))

#############################
#propensity score matching#
#############################

library(MatchIt)
set.seed(11001)
m.out <− matchit(treatment ~ covariate, data = simdata, distance = “logit”,
       method = “nearest”, caliper=.20, replace = FALSE)
summary(m.out, interactions = FALSE, standardize = TRUE)
#extract datafile (m.data) with matched cases
m.data<−match.data(object=m.out, group=“all”, distance = “distance”, weights =
       “weights”)
By(m.dadta, m.data$treatment, describe)
ps2 <− pscore(data=m.data, treatment ~ covariate,
       name.pscore=“ps”)
plot.pscore(ps2, main=“Propensity Score Distributions”, with.legend=TRUE,
       par.1=list(lty=1,lwd=2), par.0=list(lty=3,lwd=2),
       ylab =“”,ylim=c(0,5.5), xlim=c(0,1.0))

#########################
#ANOVA on Matched Data#
#########################
ANOVA2<−lm(response~treatment, data = m.data)
summary(ANOVA2, type = “III”)
cohen.d(response~treatment, data=m.data)
```

## Corresponding author

Greggory L. Keiffer can be contacted at: gregg.keiffer@gmail.com