經營學硏究 第23卷 特別號(1994年 12月)

평가자간 신뢰도 및 동의도에 관한 분석적 고찰

An Analytical Review of Interrater Reliability & Agreement

차 종 석*
김 영 배**

논문접수일 : 94. 2
게재확정일 : 94. 3

초 록

본 연구는 사회과학 방법론 상에서 중요한 지표인 신뢰도에 관하여 분석하였다. 신뢰도는 관찰자간 신뢰도, 측정도구의 신뢰도, 일반화 신뢰도로 구분된다. 본 연구에서는 관찰자간 신뢰도에 속하는 평가자간 신뢰도 및 동의도를 대상으로 한다.

평가자간 신뢰도는 분위기 연구, 리더십 연구, 직무분석 연구, 인사고과 및 성과평가 연구 등에서 자료의 객관성을 검증하는 중요한 지표로 사용되고 있다. 기존 연구들에서 다양한 평가자간 신뢰도 지표가 사용되었는데 평가자간 신뢰도와 동의도는 연구목적상 구분될 필요가 있다. 평가자간 신뢰도는 일관성(consistency)을 나타내는 것으로, 평가자들이 평가한 값들의 상관관계 정도 또는 균형적인 관계를 의미한다. 평가자간 동의도는 합의성(consensus)를 나타내는 것으로, 평가자들이 평가대상에 대하여 얼마나 똑같은 평가를 하는가를 의미한다.

지금까지 평가자간 신뢰도를 측정하는 지표 중에서 ICC가 가장 좋은 지표이고, 평가자간 동의도를 측정하는 것은 r_{wg}가 가장 합리적인 것으로 평가된다. 이들 지표를 정확하게 사용하기 위해서는 평가대상의 수, 분석단위, 단일항목/다항목, 일원/이원 분산분석 등을 고려해야 한다. 기존 연구들을 살펴본 결과, 분위기 연구의 경우 ICC(1)의 경우는 0.20 이상, ICC(2)의 경우는 0.60 이상이면 대략적으로 만족할만한 수준이다. r_{wg}의 경우는 0.8 이상이면 만족할만한 수준이다. 그러나 타 연구분야에서는 신뢰도 지표를 사용한 연구들의 부족하여 기준치를 제시하기가 어려운 실정이다.

이러한 분석결과를 바탕으로 이론적, 실무적 시사점을 제시하고, 평가에 영향을 주는 중요한 요인들과 추후 연구과제에 대하여 논의하였다.

* 한국과학기술원 경영정책학과 박사과정
** 한국과학기술원 경영정책학과 부교수

I. 序 論

사회과학 분야의 연구에 종사하는 대부분의 연구자들은 신뢰할만한 측정도구의 필요성을 인식하고 있다. 행동과학을 다루는 교과목이나 교과서 등에서 신뢰도(reliabiliy)는 중요한 주제로 다루어지고 있다(Mitchell, 1979). 그러나 행동과학 분야의 신뢰도 지표(index)에 대한 발전은 기존의 교육학, 심리학 분야에서 진행되어온 것에 비하면 일천한 수준이다.

최긴의 행동과학 분야의 연구에서는 평가자간 신뢰도(interrater reliability) 및 평가자간 동의도(interrater agreement)를 측정하는 연구들이 점점 증가하고 있다. 직무분석 연구(Cornelius, Carron, & Collins, 1979 ; Levine, Ash, & Bennett, 1980 ; McCormic, Jeanneret, & Mecham, 1972 ; Taylor, 1978), 분위기 연구(Drexler, 1977 ; Jones & James, 1979 ; James, 1982, Glick, 1985 ; Kozlowski & Hults, 1987), 리더십 연구(Hater & Bass, 1988), 업적평가 연구(Holzbach, 1978 ; Rothstein, 1990), 그리고 기타 자료의 객관성 문제를 다루는 연구(Arvey & Kvancevich, 1980 ; Greene, 1975 ; Ilgen & Fujii, 1976 ; Landy & Farr, 1980) 등에서 이러한 지표들이 주로 사용되고 있다.

특히, 조직의 분위기를 연구한 문헌(Drexler, Jr., 1977 ; James, 1982 ; Glick, 1985 ; Kozlowski & Hults, 1987)에서 이러한 신뢰도 지표가 많이 발견되고 있는데, 그 이유는 조직분위기의 이론적 단위(the unit of theory)는 개인이지만, 분석단위(the unit of analysis)는 팀, 조직 등의 상위수준이므로 변수 측정치에 대한 '합성이론(composition theory)'이 다루어져야 하기 때문이다. 즉, 개인의 분위기 점수가 팀 또는 조직단위로 합산되기 위해서는 '지각적 동의(perceptual agreement)'를 조사하여 합산할 수 있는 지를 판단해야 한다. 이러한 상황은 리더십이나 조직구조, 문화, 성과평가 등 조직행태 분야의 많은 변수들에게도 마찬가지로 적용될 수 있다.

따라서 평가자간 신뢰도 및 평가자간 동의도에 대한 개념을 정확하게 파악하고 적절하게 사용될 수 있다면 조직분위기와 같이 합산(aggregation)이 필요한 연구결과에 대한 객관성을 높일 수 있고 일반화에도 도움을 줄 수 있다(Tinsley & Weiss, 1975). 하지만 평가자간 신뢰도와 동의도를 나타낼 수 있는 측정지표가 다양할 뿐만 아니라 각기 나름대로 적용상에 한계점을 갖고 있기 때문에 연구자들이 자신의 연구에 어떤 지표를 선택 사용해야 하고 어떤

점을 유의해야 하는지 많은 혼란이 생길 수 있다. 연구자들은 자신의 연구목적과 측정변수의 속성에 따라 신뢰도와 동의도 개념을 구분해서 올바른 지표를 정확하게 사용할 필요가 있다.

본 연구의 목적은 신뢰도에 대한 개념 고찰에서 시작하여, 평가자간의 신뢰도 및 동의도에 대한 개념의 차이를 규명하고자 한다. 그리고 신뢰도와 동의도를 측정하는 여러가지 기존 지표(index)들을 구체적으로 비교분석함으로써 각 지표가 갖는 장점과 한계점을 이해하고자 한다. 마지막으로 조직행태 연구분야에서 이러한 지표를 사용한 기존 실증연구 결과들을 분석하여 각 지표가 정확하게 사용될 수 있는 기준, 상황 및 만족수준을 제시한다. 이러한 결과들은 분위기, 리더십, 조직구조 분야의 연구에 있어서 방법론 측면에서의 기여와 실무적 측면에서 인사고과 및 과제평가 등에도 많은 도움을 줄 것으로 기대된다.

Ⅱ. 신뢰도(reliability)의 개념

신뢰도(reliability)란 우리가 측정하거나 관찰한 값들이 일관적이고 동일한 결과를 보이는 정도를 나타내는 개념으로, 타당성(validity)과 더불어 측정에서 가장 중요한 지표이다. 신뢰도가 높다는 것은 측정자료가 안정되고(stability), 일관성(consistency)이 있으며, 정확(accuracy)해서 믿을 만하고(dependability), 예측가능성(predictability)이 높다는 의미이다(Kerlinger, 1964).

측정에서 신뢰도를 이론적으로 설명하고 있는 것은 분산(variance)이다. 일련의 측정치들은 우선 전체분산을 가지고 있으며, 이 중에서는 실제로 분포된 실질요소(true component)와 오차요소(error component)가 포함되어 있고 이들이 각 분산의 값을 갖는다.

Va	=	Vt	+	Ve
전체분산		실질분산		오차분산

신뢰도에 대한 조작적 정의는 '측정결과로 나타나는 전체분산에 대한 실질분산의 비율(Vt/Va)' 또는 '측정결과 나타나는 전체분산에 대한 오차분산의 비율을 1에서 제한 값(1--Ve/Va)'으로 표현된다(Kerlinger, 1964).

신뢰도에 대한 평가는 관찰자간 신뢰도(interobserver reliability), 심리학 연구에서 주로 사용되어 온 측정도구의 신뢰도(the reliability of the instrument), 연구결과의 일반화 연구(generalizability)로 크게 구분될 수 있다(Mitchell, 1979 ; Conrad & Maul, 1981).

첫째, 관찰자간 신뢰도는 행위 관찰 연구에서 주로 사용된 것으로, 같은 시점에서 같은 상황을 독립적인 관찰자들이 평가한 자료의 동의 정도를 나타내는 것이다. 관찰자들이 측정한 자료들이 객관적인 것임을 보이기 위해서 동시에 측정한 여러 명의 관찰자들간의 일치 정도를 계수로 제시한다.

둘째, 측정도구의 신뢰도는 하나의 개념을 설명하는 측정항목들간의 상관관계를 측정하는 것이다. 이는 기존의 심리학 연구에서 주로 사용되어 온 것으로 동일측정도구 2회 측정 신뢰도(test－retest reliability), 동등한 2가지 측정도구에 의한 신뢰도(alternative－form reliability) 등 여러가지 유형의 신뢰도 지표가 있다. 측정도구의 신뢰도에는 안정성(stability)과 동일성(equivalence)의 개념을 포함하고 있다(Selltiz, et al., 1959). 안정성은 같은 대상에 대하여 같은 측정도구로 한 시점에서 측정한 값과 그 후의 시점에서 측정한 값이 상관관계를 보이는 정도를 의미한다. 동일성은 같은 시점에서 같은 대상에 대하여 대등한(parallel) 측정도구로 평가한 값들간의 일치된 정도를 의미한다. 〈표1〉은 측정도구의 신뢰도 유형과 측정상황, 각 유형의 장·단점을 보여주고 있다. 동일측정도구 2회 측정 신뢰도는 측정도구의 안정성을 측정하고, 그외의 유형들은 측정도구의 동일성을 측정하는 것이다.

셋째, 일반화를 측정하는 계수(generalizability coefficient)는 연구결과의 일반화 정도를 나타내는 것이다. 이는 Cronbach의 일반화 이론(Cronbach, Gleser, Nanda, & Rajaratnam, 1972)에서 발전한 것으로써 신뢰도 연구에서 고려되는 개인차이나 측정오차 뿐만 아니라 다양한 요인들에 의한 변이들도 고려한다. 관찰자들(scorers), 동등한 측정도구(alternate forms), 다른 상황 등 여러 측면(facets)의 효과를 동시에 고려한다. 연구설계로는 팩토리얼(factorial) 디자인을 한다. 전체분산에서 각 측면(facet)의 분산정도의 비율을 추정하여 일반화 계수를 계산한다.

본 연구에서 다루고자 하는 신뢰도는 이들 세가지 형태의 신뢰도 중에서 행위 관찰연구에서 사용되어 온 관찰자간 신뢰도에 속하는 것이다. 평가대상에 대한 평가자들이 간의 일관성 및 합의성에 촛점을 둔 평가자간 신뢰도(interrater relibaility)의 다양한 지표들에 대하여 분석하고자 한다.

Ⅲ. 평가자간 신뢰도와 평가자간 동의도

평가자간 신뢰도는 "평가자들이 상호교환(interchangeable)될 수 있는 정도", 또는 "판단에 대한 동의 정도"의 개념이며, 평가자들이 평가결과에 대하여 어느정도 동의(agreement)하는가를 의미하고 있다(James, et al., 1984 ; Shrout & Fleiss, 1979 ; Bartko, 1976). 수학

〈표 1〉 신뢰도 측정의 유형별 측정 상황과 장 · 단점

신뢰도의 유형	측정 상황	장 점	단 점
동일측정도구 2회 측정 신뢰도 (test-retest reliability)	동일한 측정대상에 대하여 동일한 측정도구를 사용하여 1차 측정후, 재측정하여 두측정치간의 차이를 분석한다.	측정치의 상관관계가 높을 경우, 측정결과를 다른 상황까지 일반화 시킬 수 있다.	1차 측정에 대한 기억이나 개인적 또는 상황적인 요인의 변화로 인하여 2차 측정에 영향을 줄 수 있다.
동등한 2가지 측정도구에 의한 신뢰도 (alternate-form reliability)	같은 속성을 측정하는 대등한(parallel) 2가지 측정도구로 동일한 대상을 측정하여 이들 측정치간의 관계를 분석한다.	같은 질문에 대하여 다른 표현으로 짧은 시간내에 측정함으로써 평가자 개인적 요인이나 평가대상의 상황적 요인의 변화가 일어날 가능성이 적다.	대등한 형태의 측정도구를 완벽하게 만들기가 어렵다.
항목분할 측정치의 신뢰도 (split-half reliability)	같은 속성을 측정하는 다수의 측정항목들을 대등하게 반으로 나누어 측정함으로써 이들 두 그룹간의 관계를 분석한다.	한번의 테스트내에서 항목들의 일관성을 측정할 수 있다.	반으로 나누는 방법에 따라서 결과의 차이가 날 수 있으며, 다른 테스트와의 일관성을 보여 주지는 못한다.
내적일관성 (internal consistency reliability)	동일한 개념을 측정하기 위한 여러항목들의 상호상관관계를 조사함으로써 일관성을 분석한다.	신뢰도를 저해하는 항목을 순차적으로 제거해 나감으로써 신뢰도가 높은 항목들을 구성할 수 있다.	다른 테스트와의 일관성을 보여 주지는 못한다.

(참조) Selltiz, et al.(1959), Mitchell(1979), Conrad & Maul(1981)

적으로 앞서 말한대로 평가결과의 총분산 중에서 체계적인 분산의 비율을 나타낸다.

평가의 총분산은 무작위적 오차 분산(random error varance)과 체계적 분산(sustematic variance)로 구성된다. 무작위적 오차 분산은 측정과정에서 우연적이며, 일시적으로 나타나는 것이다. 이는 측정과정의 분위기, 통제가 어려운 상황적인 요인들, 그리고 평가자의 피곤이나 감정적인 긴장 등에 의하여 발생하는 비체계적인 오차이며 그 근거를 알기가 어려운 것이다. 체계적 분산은 측정과정에서 체계적으로 나타나는 분산이다. 평가자의 지식, 신분, 정보, 인간성 등의 요인이 영향을 준다(James, et al., 1984). 이는 다시 순수한 분산(true variance)과 체계적 오차(systematic error)로 구분되는데, 체계적 오차는 사회적으로 바람직한 응답(socially desirable response)을 하도록 하는 것과 같은 응답 편의(response bias)를 반영하고 있다. 응답 편의는 체계적 분산을 증가시켜 평가자간 신뢰도를 증폭(inflate)시키기 때문에 문제를 야기시킬 수 있다(Guion, 1965).

그러나 이러한 "평가자들이 상호교환 될 수 있는 정도" 및 "판단에 대한 동의 정도"를 나타내는 평가자간 신뢰도는 평가자간의 일관성(consistency)과 평가치의 수준(절대치)을 동시에 포함하고 있다(Lahey, 1983). 이는 광의의 평가자간 신뢰도 개념이라 할 수 있는데 초기의 연구자들은 이처럼 평가자간 신뢰도와 평가자간 동의도를 구별하지 않고 사용하였다. 그러나 최근의 연구자들은 신뢰도와 동의도가 서로 다른 측면을 나타내고 있음을 밝히면서, 이를 구분해서 사용해야 함을 주장하고 있다(James, et al., 1993 ; Kozlowski & Hattrup, 1992 ; Mitchell, 1979 ; Tinsley & Weiss, 1975).

협의의 평가자간 신뢰도는 평가자들이 평가한 값들의 관계가 균형적(proportional)인 정도를 나타내는 것이다. 한 평가치와 다른 평가치의 절대값은 다를지라도 그 상관관계가 어느정도로 일관적인가를 의미한다(Kozlowski & Hattrup, 1992 ; Tinsley & Weiss, 1975). 즉, 평가자들간 분산의 일관성에 대한 비율(Lawlis & Lu, 1972), 혹은 평가치간의 상관관계(James, et al., 1991 ; Lawlis & Lu, 1972 ; Shrout & Fleiss, 1979)를 나타내는 개념이다. 신뢰도의 값은 주로 상관관계분석과 분산분석을 통하여 계산되며, 집단내 상관관계(ICC : intraclass correlation) 분석(Shrout & Fleiss, 1975)이 대표적인 평가자간 신뢰도 지표이다.

반면에 평가자간 동의도는 평가자들이 평가대상에 대하여 절대치가 같은 평가를 하는 정도를 나타내며, 평가자들간에 상호교환이 가능한(interchangeable) 정도를 의미한다(James, et al., 1991 ; Bartko, 1976 ; Tinsley & Weiss, 1975). 평가대상에 대하여 점수로 평가를 하

는 경우, 평가자간 동의는 평가자가 서로 똑같은 수치를 부여하는 것을 의미한다. 이러한 평가자간의 신뢰도와 동의도의 정의 및 평가기법들의 차이가 〈표 2〉에 요약되어 있다.

신뢰도와 동의도의 선택은 연구목적에 따라서 구분하여 결정된다. 예를 들어, 평가 대상의 순위나 평가자들간 판단의 일관성에 관심이 있을 경우에는 평가자간 신뢰도를 측정하는 지표를 사용하여야 하고, 평가대상의 절대적인 점수 또는 평가자들간 판단의 기준치에 대해서 연구의 촛점이 있을 경우는 평가자간 동의도를 측정하는 지표를 사용하여야 한다. 두 개념을 구분하지 않고 사용할 경우 많은 연구결과들간에 혼란을 야기시킬 수 있다. 신뢰도와 동의도의 계수는 일반적으로 비슷한 값을 갖지만 꼭 그렇지만은 않다. Tinsley & Weiss(1975)의 연구는 평가치의 절대값이 다르더라도 신뢰도는 높은 경우와 신뢰도는 낮을지라도 평가치의 절대값이 비슷하여 동의도는 높은 경우를 보여주고 있다. 동의도는 낮지만 신뢰도가 높은 예는 평가자들간의 기준치는 다르지만 평가에 대해 일관된 잣대를 갖고 있는 경우에 나타나고, 반면에 동의도는 높지만 신뢰도가 낮은 예는 평가대상이 모호하여 중심화 경향이 발생함으로써 평가치는 비슷한 값을 갖지만 일관된 잣대가 없는 경우에서 발견할 수 있다.

〈표 2〉 평가자간 신로도 및 평가자간 동의도

	평가자간 신뢰도 (Consistency)	평가자간 동의도 (Consensus)
정 의	평가자들이 평가한 값의 절대치와는 상관없이, 평가치들의 상대적인 관계의 균형적인(proportional)인 정도 - 평가치들간의 분산의 일관성에 대한 비율 - 평가치들간의 상관관계의 정도 - 평가대상에 대한 순서의 유사성	평가자들 사이의 상호교환(interchangeability) 가능성 - 평가자들이 평가대상에 대하여 똑같은 평가를 하는 정도
평가 기법들	(1) 쌍대상관계수 (pairwise correlation coefficients) (2) ICC(Intraclass Correlation) (Shrout & Fleiss, 1979)	(1) 동의비율(the percetage of agreement) (2) Lawis & Lu's(1972) Chi-square and Tinsley & Weiss's(1975) T index (3) Cohen's(1968) weighted kappa (4) Finn's(1970) index (5) James's(1984) :

3-1. 평가자간 신뢰도(interrater reliability)

서열척도와 등간척도로 평가된 결과치들에서 평가자간 신뢰도를 측정하는 대표적인 지표는 평균 쌍대 상관계수(average pairwise correlation coefficient)와 ICC(intraclass correlation)이다.

평균 쌍대 상관계수는 평가자들이 평가대상에 대한 상대적인 순위들간의 유사성을 의미한다. 이는 서로 다른 평가대상에 대해서 비슷한 판단기준을 갖고 있는지를 알아보기 위한 지표라고 할 수 있다(Glick, 1985).

한편, ICC는 평가자들간의 평가가 안정적이고 일관된 정도를 나타내고 있다. ICC는 총분산 중에서 평가대상의 분산 비율을 의미하는 것(Tinsley & Weiss, 1975)으로, ANOVA 분석으로부터 쉽게 구할 수 있다. 이때 평가자간 분산이 오차항(error term)으로 포함될 경우는 일원(one-way) 분산 분석을 사용하고, 평가자간 분산이 오차항(error term)에 포함되지 않을 경우에는 이원(two-way) 분산분석을 사용한다. 순수한 평가치의 일관성(consistency)에만 관심이 있을 경우에는 평가자들간 평균 차이를 오차항에 포함하지 않는 이원 분산분석을 사용한다. 그러나 평가자가 대체될 수 있는 경우나 임의 표본추출된(random sampling) 다른 평가대상으로 일반화를 원하는 경우에는 평가자간 분산을 오차항에 포함하는 일원 분산분석을 실시한다.

---

1) ICC(1)은 ICC(*, 1)에 해당하는 모든 지표를 나타낸다.
ICC(2)은 ICC(*, K)에 해당하는 모든 지표를 나타낸다.
여기서, * : 상황 1, 2, 3
K : 2이상의 정수로 평가자 수를 나타냄
상황 1. : 평가자 모집단으로부터 임의추출된(random sampling) 평가자들 중에서, 각 평가대상은 서로다른 k명의 평가자들로부터 평가를 받는다.
(상황 1.의 모형) ⇒ $x_{ij} = u + b_j + w_{ij}$
i : 평가자($i = 1, \ldots, k$),
u : 평가대상($j = 1, \ldots, n$),
b_j : 모평균과 j 평가대상의 참값과의 차이
w_{ij} : 오차항
상황 2. : 평가자 모집단으로부터 k명의 평가자가 임의추출(random sampling)되고, k명의 평가자는 모든 n개의 평가대상을 각각 평가한다.
상황 3. : 평가자는 사전지식이 있는 관계자 k명으로 구성되고, k명의 평가자는 모든 n개의평가대상을 각각 평가한다.
(상황 2. & 상황 3.의 모형) ⇒ $x_{ij} = u + a_i + b_j + (ab)_{ij} + e_{ij}$
a_i : 모평균과 i 평가자의 평균과의 차이
$(ab)_{ij}$: i 평가자가 j 평가대상을 평가할 때 발생하는 상호작용의 정도
e_{ij} : i 평가자가 j 평가대상을 평가할 때의 평가오차

신뢰도를 측정하는 ICC는 다시 [1]ICC(1)와 ICC(2)로 구분된다. ICC(1)은 개별평가치에 대한 평가자간 신뢰도 계수를 측정하는 것으로 개별항목이 분석단위일때 사용한다. ICC(2)는 평가집단의 평균에 대한 신뢰도를 나타내는 것으로 평가집단의 합산된 점수(composite rating)가 분석단위일때 사용되며 ICC(2)는 Spearman-Brown 공식에 따라서 구해진다. 구체적인 공식내용은 〈표 3〉에 정리되어 있다.

〈표 3〉 여러가지 ICC공식

	ICC 공식		연 구 자
ICC(1) vs. ICC(2)	(one-way ANOVA) $ICC(1)=\frac{BMS-WMS}{BMS+(k-1)WMS}$ (two-way ANOVA) $ICC(1)=\frac{BMS-EMS}{BMS+(k-1)EMS}$	$ICC(2)=\frac{k[ICC(1)]}{1+(k-1)[ICC(1)]}$ ⇒ Spearman-Brown Formula	Bartko (1966), Winer (1971), Tinsley & Weiss (1975), James (1982)
상황 1. 상황 2. 상황 3.	$ICC(1,1)=\frac{BMS-WMS}{BMS+(k-1)WMS}$ $ICC(2,1)=\frac{BMS-EMS}{BMS+(k\text{-}1)EMS+k(JMS\text{-}EMS)/n}$ $ICC(3,1)=\frac{BMS-EMS}{BMS+(k-1)EMS}$	$ICC(1,k)=\frac{BMS-WMS}{BMS}$ $ICC(2,k)=\frac{BMS-EMS}{BMS+(JMS-EMS)/n}$ $ICC(3,k)=\frac{BMS-EMS}{BMS}$	Shrout & Fleiss (1979), Lahey, Downey, & Saal(1983)
사용목적	- 개인수준의 신뢰성 - 개별항목에 대한 평가자들간의 신뢰성(The reliability of a single rating)	- 조직(팀)수준의 신뢰성 - 조직별로 합산된 값에 대한 신리성(The reliability of composite rating)	

(참조) Shrout and Fleiss(1970)

BMS = mean square between targets ; WMS = mean square within targets ;
JMS = mean square between judges ; EMS = residual mean square ;
n = number of targets ; K = number of judges.

상황 1과 상황 2, 3의 차이는 일원 분산분석 혹은 이원 분석분석의 상황적 차이며, 상황 2와 상황 3은 평가자가 임의추출된 것인지, 고정된 것이지에 따라서 구별된다. ICC(*, 1)은 개별 평가치에 대한 측정의 신뢰도를 나타내며 ICC(*, k)는 k명의 평가자의 평균값에 대한 신뢰도를 나타낸다. 이는 각각 ICC(1)과 ICC(2)에 해당한다.

ICC(1)은 개별평가치에 대한 평가자간의 신뢰도에 대한 점 추정치(point estimate)이다(Winer, 1971). 일원 분산분석에 의한 높은 점수의 ICC(1)(즉, ICC(1,1)에 해당)은 본질적으로 평가대상내의 평가자간 분산이 낮은 것을 내포하고 있다(Bartko, 1976). 이는 개별항목에 대한 평가에서 평가자들간의 분산이 작아 평가자들의 합의(consensus)가 높다는 것을 의미하고 있기 때문에 평가자들간의 동의(agreement)정도의 지표로도 사용될 수 있다(James, 1982). James(1982)는 분위기 연구에서 지각적 동의(perceptual agreement)가 분위기 점수를 합산하기 위한 전제조건이기 때문에, 일원 분산분석에 의한 ICC(1)을 사용하여 개별 분위기 점수를 합산할 수 있는 지의 여부를 판단할 수 있다고 주장한다.

반면, ICC(2)는 평가집단의 평균치가 안정된 값인가의 여부를 나타낸다. 이는 서로 다른 두 그룹의 평가자들을 임의추출하여, 이들 두 그룹간의 합산된 점수(composite rating)에 대한 상관관계 정도를 구했을 때 해당되는 값이다. ICC(2)는 평가자(k)의 수가 증가하면 높아지는 특성(Guilford, 1954)을 갖고 있기 때문에 ICC(1)은 낮고 ICC(2)는 높은 경우를 생각할 수 있다. 예로 k＝300명, ICC(1)＝0.5일때 ICC(2)는 0.94가 된다(〈표3〉의 공식 참조). 이처럼 평가자 수가 많은 경우, 개별항목에 대한 합의 정도가 낮다 할지라도 평가집단에 근거한 합산된 점수의 신뢰성은 높아질 수 있다. 왜냐하면 평가자 수가 많을 경우, 평가자 개인들의 차이가 있다 하더라도 평가집단의 평균은 특정값에 수렴할 가능성이 높기 때문이다.

평균 쌍대 상관계수와 ICC를 사용하여 평가자간 신뢰도를 측정한 기존 연구들을 〈표 4〉에 요약되어 있다. 연구 분야별로, 분위기 연구에서는 이들 지표를 많이 사용하고 있는 반면 리더십 연구나 성과를 측정한 연구에서는 상대적으로 많지 않음을 알 수 있다.

ICC 지표가 수학적으로 갖고 있는 문제점은 다음 세가지로 요약될 수 있다. 첫째, 자료의 분포가 본질적으로 분산이 아주 적은 경우에는 ICC의 결과치가 무의미하게 된다. 평가자들이 평가대상에 대한 정보가 없거나, 표준화된 측정도구를 사용하여 평가대상의 특성을 제대로 반영하지 못한 경우에 평가치는 중심화 경향을 보일 가능성이 높다. 따라서 ICC를 사용하기 전에 평가대상간의 차이가 있는지를 알기 위해 두집단간의 분산차이에 대한 F-test를 먼

저 해 보아야 한다.

둘째, 다항목 측정도구의 경우 차원별로 합산된 점수에 근거한 평가자간 신뢰도 값은 개별 항목에 근거한 평가자간 신뢰도 값보다 과대 평가된 결과를 보인다. Jones, et al.(1983)은 PAQ(Position Analysis Questionnaire)를 사용하여 개별항목에 근거한 평가자간 신뢰도 값과 차원별로 합산된 점수에 근거한 평가자간 신뢰도 값을 비교하였다. 차원별로 합산된 점수에 근거한 평가자간 신뢰도 값이 높은 결과를 보여 주고 있다. 그밖의 기존연구들 중에서는 개별항목에 근거한 평가자간 신뢰도를 계산한 연구들(Bass, et al., 1975) 보다는 차원별로 합산된 점수에 근거한 평가자간 신뢰도를 계산한 연구들(Jons & James, 1979 ; Peterson, 1975 ; Zohar, 1980 ; Joyce & Slocum, 1984)이 많다. 타 연구들의 측정치와 비교할때 이점을 주의하여야 한다(〈표 4〉 참조).

〈표 4〉 평가자간 신뢰도(interrater reliability)에 관한 기존 연구들

연구논문	측정대상	평 가 자	측 정 지 표	측 정 결 과	비 고
Ilgen & Fujii (1976)	그룹리더	구성원 vs. 전문 관찰자	avergae pairwise correlation coefficient	(.06－.43) : 구성원 (.66－.96) : 전문관찰자	그룹리더의 리더십(고려, 구조주도)에 대한 쌍대(pairwise)상관관계의 평균
Curtis, Smith & Smoll (1979)	야구코치	관찰자들	구체적으로 언급되어 있지 않음	.88	야구코치의 리더십에 관한 평가자간 신뢰도
Borman (1982)	군인들의 성과	두명씩의 평가자들	pairwise correlation coefficients	.44－.92 (median＝.76)	훈련소의 군인들에 대한 교관들의 평가
Greene (1975)	종업원의 성과	동료들	Spearman-Brown Formula ICC(1, k)	.75－.89	동료들이 한 종업원에 대한 평가
Dess & Robinson (1984)	조직성과	각 조직의 최고 경영층	ICC(1, 1)	총매출액 : .87 자산회수율 : .84 총기업성과 : .84	같은 산업내의 경쟁사에 대하여 조직성과에 대한 주관적인 평가를 함

연구논문	측정대상	평가자	측정지표	측정결과	비고
Bass, et al. (1975)	과업진단 (work group)	그룹 구성원	ICC(1,k)	.134－.973 (median＝.745)	3개의 표본을 대상으로 31개 분위기 변수에 대한 신뢰도 → 총 93개 ICC(1,k)
James, et al.(1980)	한 조직의 부서들	구성원	ICC(1,k)	.512	부서간 갈등에 대한 부서별 점수에 대한 신뢰도
Jones & James (1979)	미해군의 과업환경	해군 장병들	ICC(1,1)	.01－.22 (median＝.06)	6개 차원의 심리적 분위기(psychological climate)에 대한 신뢰도
			ICC(1, k)	.55－.91 (median＝.71)	
Peterson (1975)	15개의 조직	조직 구성원	ICC(1,1)	extrinsic motivation .563 organizational style .615 intrinsic motivation .519 leadership style .288	4개 차원의 조직분위기에 대한 신뢰도
			ICC(1,k)	extrinsic motivation .989 organizational style .991 intrinsic motivation .987 leadership style .965	
Zohar (1980)	20 공장	각 공장에 속한 근로자 (각각 20명)	ICC(1,1)	.721	안전에 대한 조직분위기(organional climate for safety)신뢰도
			ICC(1,k)	.981	
James (1982)	－	－	ICC(1,1)	.00－.50 (median＝.12)	13개 기존연구의 신뢰도 계수를 사용하여 ICC(1,1)을 다시 계산함
Joyce & Slocum (1984)	3 공장	각 공장의 십장들	ICC(1,1)	4개 : .56－.79 10개 : .24－.42 2개 : .10, .14 2개 : .00, .06	3개 공장에서 6개 분위기 차원에 대한 신뢰도 (총 18개)

셋째, 같은 자료를 사용하여 계산된 ICC(1)과 ICC(2)를 비교하여 보면 ICC(2)가 항상 높게 나타나고 있다(〈표 4〉 참조). 수학적 공식으로 ICC(2)는 ICC(1)보다 높거나 같은 값을 갖고 있다(Bartko, 1976). 〈표 4〉에 나타난 Jones & James(1979)의 연구결과는 ICC(1)과 ICC(2)를 함께 보여 주고 있다. ICC(1)의 중위치는 0.06이고, ICC(2)의 중위치는 0.71로 나타나 많은 차이를 보이고 있다. 그외 연구들(Peterson, 1975 ; Zohar, 1980)에서도 ICC(2)가 높게 나타나고 있다. 따라서 측정의 목적에 맞는 적절한 지표를 선택하여 사용하여야 한다.

3-2. 평가자간 동의도(interrater agreement)

동의도(agreement) 개념을 반영한 초기의 지표들은 "동의의 퍼센트나 비율"을 의미한다(Cohen, 1968 ; Lawlis & Lu, 1972 ; Lu, 1971). 이들 지표들은 쉽게 계산할수 있고 간편하게 이용할 수 있지만, 평가치의 우연한 동의(chance agreemant)와 사용되는 척도의 선택점(categories)의 수에 대한 고려를 하지 못하는 약점을 지니고 있다. 동의도에 대하여 '예, 아니오'의 이분법이 적용되기 때문에 서열척도나 등간척도의 경우에는 본질적인 문제를 내포한다(Kozlowski & Hattrup, 1992).

평가치의 절대값에 의한 이분법적인 동의의 한계를 극복하기 위해서는 동의의 범위를 설정할 필요가 있다. Lawlis & Lu(1972)의 X^2-test와 Tinsley & Weiss(1975)의 T 지표는 평가치가 1 또는 2점 차이가 나는 것은 동의하는 것으로 간주하여 계산하고 있다. 일치/불일치의 절대값에 의한 이분법의 적용보다는 융통성을 지니고 있는 것으로 연구목적에 따라서 적절하게 정해져야 한다. 또한 Cohen(1968)이 사용한 weighted kappa는 불일치의 정도에 가중치을 두어 계산한다. 그러나 이들 두 지표들도 평가치의 우연한 동의를 고려하지 못하는 약점을 지니고 있다.

동의도의 지표에 우연한 동의가 고려되지 않을 경우에는 동의도의 값은 과도하게 계산되어진다. Finn(1970)의 연구와 이를 기반으로 한 연구들(Tinsley & Weiss, 1975 ; James, et al., 1984 ; Kozlowski & Hults, 1987)은 완전히 무작위적으로 응답할(completely random responding) 경우를 동의도의 지표에 반영하고 있다. 대개의 지표들은 응답자의 무작위적 응답형태를 정방형 분포(rectangular distribution)로 간주하여 반영하고 있다. 무작위적으로 응답할 경우, 선택점들이 모두 같은 확률을 갖고 있다는 논리이다.

Finn(1970)의 지표(r)에서 고려한 정방형 분포는 우연한 동의를 고려한 연구들의 표준점이 되었다.

$$r = 1.0 - \frac{\text{observed variance}}{\text{chance varinace}(s_c^2)}$$

$$(s_c^2) = \frac{k^2 - 1}{12}$$

(s_c^2) : The expected or chance variance　　k = the number of points

그는 관찰된 분산(observed variance)이 무작위 분산(random variance)보다 작은 정도가 동의(agreement)의 정도라고 설정하고, 1에서 관찰된 분산에 대한 무작위 분산의 비율을 뺀 값을 동의도의 계수(r)로 사용한다. 그러나 Finn(1970)의 지표(r)에서도 응답 편의(responce bias)가 존재할 경우 정방형 분포로 표현된 무작위 분산은 과도하게 산정된 결과를 초래하여 동의도의 계수가 과도하게 추정되는(overestimate) 약점을 갖고 있다.

James, et al.(1984)의 연구는 Finn(1970) 개념을 기반으로 하여 분위기에 대한 합산 모형(composite model)을 위한 합의(consensus)를 측정하는 지표를 개발하였다. 그룹내 신뢰도(within-group reliability)를 나타내는 이 지표는 한 평가대상에 대하여 개별항목에 대한 편가자간의 동의도($r_{wg(1)}$)와 다항목에 대한 평가자간 동의도($r_{wg(1)}$)로 구분된다.

$$r_{wg(1)} = 1 - (S_{xj}^2 / Q_{EU}^2)$$

$$r_{wg(J)} + \frac{J\,[\,1 - (\overline{S_{xj}^2} / Q_{EU}^2)}{J\,[\,1 - (\overline{S_{xj}^2} / Q_{EU}^2) + (\overline{S_{xj}^2} / Q_{EU}^2)\,]}$$

J : 다항목의 갯수
S_{xj}^2 : xj에 대한 관찰된 분산
Q_{EU}^2 : 모든 평가가 무작위적(random)으로 일어난다고 할 경우의 xj에 대한 분산
$\overline{S_{xj}^2}$: 관찰된 분산의 J개의 항목에 대한 평균

〈표 5〉는 평가자간 동의도 지표를 사용한 기존 연구들을 보여주고 있다. 〈표 5〉에 나타나듯이 평가자간 동의도 개념을 측정하기 위한 r_{wg}지표를 사용한 연구결과들은 아직 많지 않다.

최근 일관성(consistency)과 합의성(consensus) 개념을 구분할 필요성(James, et al., 1984 ; Kozlowski & Hattrup, 1992 ; James, et al., 1993)이 대두되면서 r_{wg} 지표를 사용하기 시작했기 때문이다.

r_{wg} 지표에 대한 수학적인 현상 및 문제점은 두가지로 요약될 수 있다. 첫째, 다항목에 대한 평가자간 동의도($r_{wg(J)}$)가 개별항목에 대한 평가자간의 동의도($r_{wg(1)}$)보다 높게 나타난다. $r_{wg(J)}$의 경우는 차원별 합산된 점수에 근거한 평가자가 동의도이다. 이 경우는 하나의 개념을 측정하는 여러항목을 평균함으로써 측정오차의 영향을 줄일 수 있다. 그 결과 $r_{wg(J)}$는 $r_{wg(1)}$보다 높은 결과를 나타낸다. 기존연구들에서는 중요 변수들이 다항목이었기 때문에 주로 $r_{wg(J)}$를 사용하였다.

둘째, r_{wg}의 지표도 Finn(1970)의 지표처럼 무작위 분산(ramdom variance)을 가정하여 개발된 것이다. 이는 평가자들의 응답분포가 특별한 분포를 갖고 있을 때는 왜곡된 결과를 초래할 수 있다. James, et al.,(1984)는 가능한 응답 바이어스(예, central tendency, leniency, social desirability)를 고려한 기대분포(expected distribution)를 설정하여 집단내 동의도를 평가하는 방법을 제안하고 있다. 응답 편의가 없을 때의 최대 기대분포와 응답 편의가 있을 때의 최소 기대분포를 설정하여 특정상황에서 r_{wg}의 범위(range)를 제시하였다. 이때 J(평가항목의 수)가 증가함으로써 r_{wg}의 범위가 좁아지는 특성을 갖고 있다. 실제적으로 관찰한 분산에는 응답 편의가 들어 있는지의 여부를 알 수 없기 때문에 동의도 수준의 범위를 제시하는 것이 현실적이라 할 수 있다. Kozlowski & Hults(1987)의 연구에서는 무작위 분산을 가정했을 때는 0.81−0.96이지만, 사회적으로 바람직한 응답을 고려한 기대분포(skewed distribution)를 가정했을 때는 0.70−0.95로 약간 낮게 나왔다(〈표 5〉 참조).

Ⅳ. 평가자간 신뢰도 및 동의도의 사용지침

평가자간 신뢰도 및 동의도에 관한 많은 연구들은 여러가지의 신뢰도 및 동의도에 대한 공식을 소개하고 있다. 그러나 이들 공식들은 제한된 상황속에서만 사용될 수 있다(Shrout & Fleiss, 1979 ; Jones, et al., 1983 ; Kozlowski & Hattrup, 1992). 각 상황과 여러지표들의 장·단점에 대한 분석을 통하여 특정목적에 타당한 지표를 사용하기 위한 지침들을 설정할 수 있다.

〈표 5〉 평가자간 동의도(interrater agreement)에 관한 기존 연구들

연구논문	측정대상	평 가 자	측 정 지 표	측 정 결 과	비 고
Schneider & Bowen (1985)	은행의 지점 (branch)	은행원	$r_{wg(J)}$	(인적자원관리제도) supervision .86, status .69, career facilitation .88, socilization .73, work facilitation .92	여러 부서들 중에서 구성원이 가장 많은 부서(8명)에서만 인적자원관리의 5가지 차원에 대한 Agreement를 계산함
Hater & Bass (1988)	그룹 리더	그룹 구성원	$r_{wg(J)}$	(MLQ : Multifactor Leadership Questionnaire) 카리스마 .89 개인적고려 .89, 지적자극 .90, 상황적 보상 .72, 예외적 능동관리 .71, 예외적 수동관리 .67	MLQ의 6차원에 대하여 58명의 그룹리더 각각에 대한 Agreement를 구하여 평균한 값임.
Kozlowski & Hutls (1987)	10개의 조직	조직 구성원	$r_{wg(J)}$	.81－.96 .70－.95 (skewed distribution을 가정)	기술적 개선분위기에 대한 6차원에 대하여 10개 조직 각각의 $r_{wg(J)}$ 계산(사회적 바람직한 응답을 가정하여 기대분포를 skewed distribution을 전제로 함)
Sackett & Wilson (1982)	평가 위원회	평가 위원	Percent of disagreement	4.6－37.4 %	평가과정에 영향을 주는 요인들에 대한 disagreement의 퍼센트를 측정함

평가자간 신뢰도에 대한 각 지표들은 한계를 지니고 있음에도 불구하고 현재까지 가장 좋은 것으

로 받아들여지고 있는 것은 집단내 신뢰도(r_{wg})와 ICC이다(James, 1982; James, et al., 1984, 1993; Glick, 1985; Kozlowski & Hattrup, 1992). 이들 지표들을 적절하게 사용할 수 있는 사용기준, 구체적인 상황, 만족수준을 살펴봄으로써 평가자간 신뢰도 및 동의도를 측정하고자 하는 연구들에게 이론적, 실증적인 도움을 줄 수 있다.

4-1. ICC와 r_{wg}의 사용기준

여러가지 측정지표 중에서 측정목적에 적합한 지표를 선택하기 위해서 다음과 같은 사항들이 고려되어야 한다.

첫째, 신뢰도의 측정목적이 동의도(agreement)에 대한 것인지, 신뢰도(reliability)에 대한 것인지를 분명히 파악해야 한다. 즉, 측정목적이 평가자들의 평가수준(절대치)이 같은 정도든지, 평가자들간의 상호교환 가능성 정도를 평가하고자 할때는 동의도를 측정하는 지표를 사용한다. 그러나 측정목적이 평가자들이 평가한 값들간의 균형적인 정도 및 평가자들간의 일관성의 정도를 평가할 때에는 신뢰도를 측정하는 지표를 사용한다.

둘째, 평가항목이 단일항목인지, 다항목 척도인지를 파악한다. ICC의 경우, 다항목 척도에서 차원별 합산된 점수에 근거한 ICC값은 개별항목에 근거한 ICC값보다 과대 계산된다는 사실에 주의하여 개별항목에 근거한 ICC값을 제시할 필요성이 있다. r_{wg}의 경우, 단일 항목일 때에는 $r_{wg(1)}$를 사용하고, 다항목일때에는 $r_{wg(J)}$를 사용한다.

셋째, ICC를 사용할 경우, 평가상황이 일원(one-way) 분산분석인지, 이원(two-way) 분산분석인지를 결정한다. 평가자들이 임의추출되어 평가자들이 일반성을 가져야할 경우에는 평가자간 분산이 오차항에 포함되는 일원 분산분석을 사용한다. 그러나 평가자간의 순수한 일관성에만 관심이 있을 경우에는 평가자들간의 분산이 오차항에 포함되지 않는 이원 분산분석을 사용한다.

넷째, ICC의 경우, 개인별 점수에 대한 평가자간 신뢰도인지, 평가집단의 합산된 점수에 대한 신뢰도인지를 분명하게 한다. 개인별 점수에 대한 평가자간 신뢰도를 측정하고자할 경우에는 개인수준의 신뢰도 지표인 ICC(1)을 사용하고, 평가집단의 합산된 점수에 대한 평가자간 신뢰도를 측정하고자할 경우에는 팀(조직)수준의 신뢰도 지표인 ICC(2)를 사용한다.

다섯째, ICC의 경우, 평가자들이 고정된(fixed) 집단인지, 임의추출된(random sampling)

집단인지를 파악한다. 평가자들이 특정집단인 경우에는 ICC(2,1), ICC(3,1), ICC(2,k), ICC(3,k)을 사용하고 임의추출된 집단인 경우에는 ICC(1,1), ICC(1,k)를 사용할 수 있다.

여섯째, r_{wg}를 사용할 경우, 평가대상(target)이 하나인지 여러개를 대상으로 하는지를 고려한다. 평가대상이 여러개일 경우, 각 평가대상에 대한 r_{wg} 값을 구하여 모든 평가대상에 대한 r_{wg}를 평균하여 r_{wg} 값을 제시한다.

〈표 6〉은 ICC와 r_{wg}를 평가대상의 수와 신뢰도 분석의 목적에 따라서 구분하고 있다. 평가대상은 한 개일때와 n 개일때로 구분되고, 분석의 목적은 한 평가항목에 대한 동의도 및 신뢰도, 다항목을 합산한 차원별 점수에 대한 동의도, 그리고 평가집단의 합산된 점수에 대한 신뢰도에 따라서 분류되어진다.

〈표 6〉 ICC와 r_{wg}의 사용기준

평가 대상(Target)의 수	신뢰도(Reliability) 분석의 목적		
	한 평가항목에 대한 동의도 및 신뢰도	대등한(parallel) J개의 항목들(다항목)을 합산한 값에 대한 동의도	평가집단의 합산된 점수에 대한 신뢰도
1 개	Within-group reliability ($r_{wg(1)}$)	Within-group reliability ($r_{wg(J)}$)	–
n 개	ICC(1)	–	ICC(2)

4－2. ICC와 r_{wg}의 적용상황

앞에서 제시한 사용기준들에 근거하여 ICC와 r_{wg}의 구체적인 적용상황를 살펴보기로 하자. 각 적용상황에 대한 충분한 이해는 타 분야에서 본 지표를 사용할때 많은 참조가 될 것으로 기대한다.

먼저, ICC지표에서, ICC(1)는 평가대상(target)이 여러 개이고 개인수준에서 평가자들간의 신뢰도를 측정할때 사용하고, ICC(2)는 평가대상이 여러 개이고 조직(팀)수준의 값을 구하기 위해 평가자들의 값을 평균(composite rating)할 때, 평균값에 대한 평가자간 신뢰도를 측정할때 사용한다.

그리고 r_{wg} 지표에서 $r_{wg(1)}$ 는 평가대상이 하나이고 단일항목만으로 평가를 할때 평가자 집단내의 평가자들간의 측정에 대한 동의도를 측정할때 사용하고, $r_{wg(J)}$ 는 평가대상이 하나이고 다항목 측정도구로 평가를 할 때, 평가자 집단내의 평가자들간의 측정에 대한 동의도를 측정할때 사용한다.

각 지표에 대한 구체적인 적용상황은 〈표 7〉에서 제시하고 있다.

〈표 7〉 ICC와 r_{wg}에 대한 구체적인 적용상황

(1) ICC(1)	평가대상(target)은 여러 개이고 개인수준에서 평가자들간의 신뢰도(interrater reliability)를 측정함.
ICC(1)의 적용예시	① 각 연구개발 프로젝트팀의 구성원들(k명)이 소속 팀의 '상업적 성공에 대한 기여도'라는 단일항목으로 측정한 결과의 신뢰도를 평가한다. → ICC(1,1) ② 한 조직에서 무작위적으로 추출된 k명이 각 연구개발 프로젝트팀에 대하여 '응집력의 정도'라는 단일항목으로 측정한 결과의 신뢰도를 평가한다. → ICC(2,1) ③ 고정된(fixed) 특정 연구개발관리 전문가 집단(panel)이 각 프로젝트팀에 대하여 '응집력의 정도'라는 단일항목으로 측정한 결과의 신뢰도를 평가한다. → ICC(3,1)
(2) ICC(2)	평가대상(target)은 여러 개이고 조직(팀)수준의 값을 구하기 위해 평가자들의 값을 평균(composite rating)할때, 평균값에 대한 신뢰도(reliability)를 측정함.
ICC(2)의 적용예시	① 개별 조직에서 임의추출된 구성원들(k명)이 해당 조직의 분위기를 평가하여 구성원들의 평균값에 대한 신뢰도를 평가한다. → ICC(1,k) ② 임의 추출된 일반인들(k명)이 국내 여러 조직의 분위기를 평가하여 개별조직에 대한 평균값의 신뢰도를 평가한다. → ICC(2,k) ③ 문화/분위기 전문가인 교수팀(k명)이 국내 여러 조직의 분위기를 평가하여 개별조직에 대한 평균값의 신뢰도를 평가한다. → ICC(3,k)
(3) $r_{wg(1)}$	평가대상(target)이 하나이고 단일항목만으로 평가를 할때, 평가자 집단내의 평가자들간의 측정에 대한 동의도(agreement)를 측정함.
$r_{wg(1)}$의 적용예시	① 학회지의 원고 게재 여부에 대하여 '아이디어의 참신성'이라는 단일항목만으로 평가자들이 평가한다. ② 연구소에서 우수 연구개발 프로젝트를 선정하는데 있어서, '상업적 성공정도'라는 단일항목만으로 평가자들이 평가한다.
(4) $r_{wg(1)}$	평가대상(target)이 하나이고 다항목 측정도구로 평가를 할 때, 평가자 집단내의 평가자들간의 측정에 대한 동의도(agreement)의 정도를 측정함.
$r_{wg(J)}$의 적용예시	① 학회지의 원고 게재 여부에 대하여 '논문의 체계성'이라는 차원에 대하여 다항목 척도로 평가자들이 평가한다. ② 연구소에서 우수 연구개발 프로젝트를 선정하는데 있어서, '프로젝트의 성과'라는 차원에 대하여 다항목 척도로 평가자들이 평가한다. ③ 조직 구성원들이 다항목 척도로 이루어진 "자율성"이라는 분위기 차원을 평가한다 (perceptual agreement를 증명함).

4--3. ICC와 r_{wg}의 만족수준

이상에서 제시된 여섯가지의 기준들과 각 상황들에 대한 충분한 이해는 신뢰도 및 동의도 지표 사용의 정확성을 높여줄 수 있다. 그러나 정확한 지표를 사용하여 그 점수가 나왔을 때, 그 결과치가 어느 정도 만족할만한 수준인지에 대한 문제가 아직 남아 있다. 연구결과로 산출된 신뢰도의 값이 어느정도 유의한가를 검증할 수 있는 통계적 방법이 아직까지 존재하지 않기 때문에 같은 분야의 기존 연구들과 비교함으로써 개략적으로 파악할 수밖에 없다.

Glick(1985)은 조직분위기에 대한 지표로 평가집단의 평균값(조직별 점수)에 대한 신뢰도, 개별항목에 근거한 평가자간 신뢰도, 차원별 합산된 점수에 근거한 평가자간 신뢰도가 있는데, 이들 3개의 지표가 적어도 0.6은 넘어야 한다고 주장한다. 하지만 기존연구들에서 평가집단의 평균값에 대한 조직수준의 신뢰도 점수는 비교적 높게 나오지만 개인수준의 신뢰도를 측정하는 ICC(1)의 경우는 높지 않다.

James(1982)는 기존에 발표된 13개의 분위기 분야의 연구결과들에서 개인수준의 신뢰도를 측정하는 ICC(1,1)을 조사하였다. 그 결과로 범위(range)가 0.00－0.50이었고 중위치는 0.13이었다. 그러나 13개의 연구들 중에서 연구설계의 타당성이 비교적 높은 연구들(Peterson, 1975 ; Zohar, 1980)의 경우에는 0.288, 0.519, 0.563, 0.615, 0.721로 높게 나타나고 있다.

신뢰도 값의 만족할만한 기준은 연구분야에 따라서 차이가 난다. 기존 연구들을 살펴본 결과, 분위기 연구의 경우는 대략적으로 ICC(1)의 경우는 0.20 이상, ICC(2)의 경우는 0.60 이상이면 만족할 만한 수준이라 할 수 있다. 타 분야에서는 ICC가 사용된 연구들이 많지 않기 때문에 한 두 개의 기존 연구들을 참고하여 비교해야 하겠다. r_{wg}의 경우에는 0.8 이상이면 대체적으로 만족할 만한 수준이라고 할 수 있다(〈표 4〉, 〈표 5〉 참조).

V. 요약 및 결론

연구방법론 상에서 변수들간의 관계를 파악하기 전에 해당변수의 측정값들에 대한 신뢰도 분석이 선행되어야 한다. 신뢰도가 낮은 경우에는 변수들간의 관계에 대한 해석이 무의미하

다. 본 연구에서는 전통적인 신뢰도의 개념을 살펴보고, 평가자간 신뢰도(interrater reliability)와 평가자간 동의도(interrater agreement)를 구분하여 그 차이점 및 해당 지표들에 대한 의미를 규명하였다.

초기의 평가자간 신뢰도는 일관성(consistency)과 합의성(consensus) 개념을 모두 포함하고 있었다. 그러나 순수한 일관성만을 측정하고자 할 때는 평가자간 평균 차이가 오차항에 포함되지 않는 이원 분산분석에서 ICC를 계산함으로써 구할 수 있다. 최근의 분위기 연구에서 지각적 합의(perceptual agreement)를 측정하기 위해서 평가자간의 동의도에 촛점을 둔 지표가 나타났다.

지금까지의 평가자간 동의도에 대한 측정지표로는 r_{wg}가 가장 합리적이고, 평가자간 신뢰도는 ICC가 가장 좋은 측정지표이다. 이들 지표(index)들을 정확하게 사용하기 위해서는 평가대상의 수, 분석단위, 단일항목 혹은 다항목, 일원 혹은 이원 분산분석 등을 고려하여 선택하여야 한다.

그러나, 이들 지표들도 아직 완벽하지 않고, 문제점을 내포하고 있다. ICC의 경우에는 평가대상간에 유의한 차이가 있을때만 의미가 있다. 평가대상간의 차이가 극미하고 평가대상내의 분산과 오차분산이 상대적으로 클 경우, ICC는 실제적인 한계(0 − 1)를 벗어날 수 있다. 따라서 연구자나 실무자들은 ICC를 계산하기 전에 평가대상간의 차이검증(F-test)을 하여야 한다(Lahey, et al., 1983). 또한 ICC는 평가자들간의 합의가 높으면서 평가대상간의 차이가 별로 없을 때에는 낮게 나타나는 약점이 있다(Tinsley & Weiss, 1975 ; Kozlowski, 1992). 이러한 문제점으로 ICC는 순수한 평가자간 동의도를 측정하는데는 부적합하다.

합의성 개념을 측정하기 위해서는 집단내 동의도 지표인 r_{wg}를 사용하는 것이 타당하다. r_{wg}지표의 문제점은 측정치의 우연에 의한 기대분포(expected distribution)를 산정할 때, 정확한 기대분포를 결정하는 좋은 방안이 없다는 점이다. 평가자들이 각자 독특한 응답 바이어스를 갖고 있을 때, r_{wg}는 이를 고려하지 못하고 있다(Lahey, et al., 1983). 또한 r_{wg}는 평가자의 수가 작을 때는 평가자간 동의도가 과소평가된다. James 등(1984)은 대체로 10명의 평가자가 적당하고 제시하고 있다.

평가자간 신뢰도 및 동의도 지표가 문제점을 내포하고 있지만, 타 연구의 이론적 측면에 많은 기여를 하고 있다. 리더십 연구, 분위기 연구 등에서 이론적 단위는 개인이지만 분석단위는 팀(조직)수준일때 평가자간 신뢰도 및 동의도를 평가함으로써 팀(조직)수준의 분석을 타

당하게 한다.

리더십 연구의 경우, 개인적인 리더십과 팀수준의 리더십으로 구분하고 있다. 팀수준의 리더십(average leadership style)을 측정할때 구성원들의 리더십에 대한 지각점수를 합산하여야 한다. 팀수준의 리더십은 팀구성원들 모두와 상호작용하는 리더의 전반적인 행태이기 때문에 팀구성원들의 지각은 유사해야 한다. 이러한 맥락에서 팀수준의 리더십을 측정하는 항목들은 먼저 구성원들간 동의도를 평가한 이후에 팀수준의 값을 구해야 한다.

분위기 연구에서도 팀(조직)분위기를 측정하고자 할때, 먼저 구성원들의 분위기 지각에 대한 합의가 선행되어야 한다. 팀(조직)분위기는 팀(조직) 특성을 반영하는 변수이고, 해당 팀(조직)의 구성원들은 그들간의 상호작용으로 인하여 주위의 현상을 비슷하게 해석 및 이해하여 유사한 의미(meanings)를 부여한다는 이론에 근거를 두고 있다(Prichard, 1973). 따라서 팀(조직)분위기 점수를 구하기 위해서는 구성원들의 합의(consensus)를 평가하는 평가자간 동의도 측정이 우선되어야 한다.

실무적으로 평가자간 신뢰도 및 동의도는 인사고과, 과제평가 등에서 중요한 주제이다. 인사고과의 경우, 평가자가 누구인가가 평가결과에 중요한 영향을 미친다. 평가자 유형은 자기, 상급자, 하급자, 전문가 등으로 구분할 수 있다. 다양한 평가자 유형들간 평가의 상관관계는 높지 않고(Landy & Farr, 1980), 대체적으로 동료나 자신에 의한 평가가 상급자에 의한 평가보다 후한 것으로 나타나고 있다(Zedeck, et al., 1974). 각 평가자 유형에 따라서 평가차이가 나는 것은 개인적인 친분이나 정보, 그리고 업적에 대한 각기 다른 시각을 갖고 있기 때문으로 파악된다. 따라서 어떤 유형의 평가자가 보다 타당하다고 말하기는 어렵지만, 정확한 평가를 위해서 피평가자의 업무에 대한 지식이 많고, 평가경험이 풍부한 사람이 평가하여야 한다.

과제평가의 경우, 먼저 과제의 특성을 잘 반영할 수 있는 평가항목의 개발이 있어야 한다. 평가항목이 특정과제의 특성만을 반영한다든지, 일률적인 평가기준을 적용하는 경우는 부적절하다. 또한 과제평가는 평가목적(선정, 진도, 최종, 사후 평가)에 따라서 중요한 요인을 고려한 평가항목이 구성되어야 한다. 과제평가에서는 평가자간 동의도보다 평가자간 신뢰도가 더 중요하다. 왜냐하면 평가자들은 자신의 기대수준에 따라서 기준치의 차이를 가져올 수 있기 때문이다. 그러나 평가자들이 과제들을 비교하는 잣대는 일관성을 유지해야 한다. 국내 연구소는 대개 평가위원회가 구성되어 있다. 이들 평가자들이 정확한 평가를 하기 위해서는 평

가과제에 대한 전문성, 평가결과에 대한 책임성, 평가기준에 대한 일관성을 갖고 있어야 한다.

정확한 평가를 위해서는 적절한 평가기법을 사용하는 것 뿐만 아니라 평가에 영향을 주는 요인들을 균형되게 유지할 수 있어야 한다. 평가자간 신뢰도 및 동의도의 값은 평가대상, 평가척도, 평가자, 그리고 평가상황의 함수로 표현될 수 있다(Tinsley & Weiss, 1975; Padgett & IIgen, 1987; Landy & Farr, 1980).

평가자간 신뢰도 및 동의도 = F (평가대상, 평가자, 평가척도, 평가상황)

이들 영향요인들에 대한 정확한 정보와 평가를 할 수 있는 상황마련이 중요하다. 첫째, 평가대상은 평가하고자 하는 모집단을 대표하는 대상이어야 한다. 둘째, 평가자는 평가경험이 풍부하고 평가대상에 대한 많은 정보를 갖고 있는 사람이어야 한다. 셋째, 평가척도의 각 항목들은 질문사항을 분명하게 하여 응답자들이 각자 다른 해석을 하지 않도록 한다. 넷째, 평가상황은 평가자 훈련을 통하여 오류(error)를 줄일 수 있도록 한다.

본 연구에서 다룬 신뢰도 지표가 행동과학의 방법론상에 기여를 하기 위해서는 앞으로 다음과 같은 점들에 대한 발전이 필요하다. 첫째, 아직까지 각 지표에 대한 만족할만 한 수준에 대해 엄격한 기준을 제시하지 못하였다. 이를 위해 먼저 수리적 통계분석에 의해서 신뢰도와 동의도 값의 유의성에 대한 가설검증을 할 수 있는 통계량이 개발되어야 한다. 둘째, 연구분야별 신뢰도지표의 실증결과치에 대한 체계적인 분석을 통하여 연구분야별로 상대적인 기준을 제시해야 한다. 셋째, 순수한 동의도를 측정하는 지표의 경우, 현재는 평가대상이 한 개인 경우만을 고려하고 있다. 앞으로 여러 평가대상을 고려한 동의도 측정지표의 개발이 필요하다.

〈참 고 문 헌〉

1. Alginal, J., "Comment on Bartko's On Various Intraclass Correlation Reliability Coefficients," *Psychological Bulletin,* Vol.85, No.1, pp.135-138(1978).
2. Arvey, R.D., & J.M. Ivancevich, "Punishment in Organizations : A review, Proposition, and Research Suggestions," *Academy of Management Review,* Vol.5, pp. 123-132(1980).
3. Bartko, J.J., "On Various Intraclass Correlation Reliability Coefficients," *Psychological Bulletin,* Vol.83, No.5, pp.762-765(1976).
4. Bartko, J.J., "Reply to Algina," *Psychological Bulletin,* Vol.85, pp.139-140(1978).
5. Bass, B.M., E.R. Valenzi, D.L. Farrow, & R.J. Solomon, "Management Styles Associated with Organizational Task, Personal, and Interpersonal Contingencies," *Journal of Applied Psychology,* Vol.60, pp.720-729(1975).
6. Borman, W.C., "Validity of Behavioral Assessment for Predicting Military Recruiter Performance," *Journal of Applied Psychology,* Vol.67, pp.3-9(1982)
7. Cohen, J, "Weighted kappa : Nominal Scale Agreement with Provision for Scaled Disagreement and Partical Credit," *Psychological Bulletin,* Vol.70, pp.213-220(1968).
8. Conrad, E., & T.Maul, *Introduction to Experimental Psychology,* New York : John Wiley & Sons(1981).
9. Cornelius, E.T., T.J.Carron, & M.N.Collins, "Job Analysis Models and Job Classification," *Personnel Psychology,* Vol.32, pp.693-708(1978).
10. Cronbach, L.T., G.C.Gleser, H.Nanda, & Rajaratnam, *The dependability of behavioral measurements : Theory of generalizability for scores and profiles,* New York : Wiley, 1972.
11. Curtis, B., R.E.Smith, & F.L.Smoll, "Scrutinizing the Skipper : A Study of Leadership Behaviors in the Dugout," *Journal of Applied Psychology,* Vol.64, pp.391-400 (1979).
12. Dess, G.G., R.B.Robinson, Jr., "Measuring Organizational Performance in the Ab-

sence of Objective Measures : The Case of the Privately-held Firm and Conglomerate Business Unit," *Strategic Marnagement Journal*, Vol.5, pp.265-273(1984).

13. Drexler, J.A., "Organizational Climate : Its Homogeneity within Organization," *Journal of Applied Psychology*, Vol.62, pp.38-42(1977).
14. Finn, R.H., "A Note on Estimating the Reliability of Categorical Data," *Educational and Psychological Measurement*, Vol30, pp.71-76(1970).
15. Glick, W.H., "Conceptualizing and Measuring Organizational and Psychological Climate: Pitfalls in Mulilevel Research," *Academy of Management Review*, Vol.10, pp. 601-616(1985).
16. Greene, C.N., "The Reciprocal Nature of Influence Between Leader and Subordinate," *Journal of Applied Psychology*, Vol.60, pp.187-193(1975).
17. Guion, R.M., *Personnel Testing*, New York : Mcgraw-Hill, 1985
18. Hater, J.J., & B.M., Bass, "Superiors' Evaluations and Subordinates' Perceptions of Transformational and Transactional Leadership," *Journal of Applied Psychology*, Vol.73, No.4, pp.695-702(1988).
19. Holzbach, R.L., "Rater Bias in Performance Ratings : Superior, Self-, and Peer Ratings," *Journal of Applied Psychology*, Vol.63, No.5, pp.579-588(1978).
20. Ilgen, D.R.,& D.S.Fujii, "An Investigation of the Validity of Leader Behavior Descriptions Obtained From Subordinates," *Journal of Applied Psychology*, Vol.61, pp.642-651(1976).
21. James, L.R., "Aggregation Bias in Estimated of Perceptual Agreement," *Journal of Applied Psychology*, Vol.67, No.2, pp.219-229(1982).
22. James, L.R., R.G.Demaree, & J.J.Hater, "A Statistical Rationale for Relating Situational Variables and Individual Differences," *Organizational Behavior and Human Performance*, Vol.25, pp.354-364(1980).
23. James, L.R., R.G. Demaree, & G. Wolf "Estimating Within-Group Interrater Reliability With and Without Response Bias," *Journal of Applied Psychology*, Vol.69, No.1, pp.85-98(1984).

24. James, L.R., R.G. Demaree, & G.Wolf, "r_{wg} : An Assessnent of Within-Group Interrater Agreement," *Journal of Applied Psychology,* Vol.78, No.2, pp.306-309 (1993).

25. Jones, A.P., & L.R.James, "Psychological Climate : Dimensions and Relationships of Individual and Aggregated Work Environment Perceptions," *Organizational Behavior and Human Performance,* Vol.23, pp.201-250(1979).

26. Joyce, W.F., & J.W.Slocum, Jr., "Collective Climate: Agreement as a Basis for Defining Aggregate Climate in Organizations," *Academy of Management Journal,* Vol.27, pp.721-742(1984).

27. Kavanagh, M.J., "Issues in Managerial Performance : Multitrait-Multimethod Analyses of Rating," *Psychological Bulletin,* Vol.75, No.1, pp.34-49(1971).

28. Kerlinger, F.N., *Foundations of Behavioral Research(2nd ed.),* Toronto : Holt, Rinehart and Winston(1973).

29. Kozlowski, S.W., & B.M.Hults, "An Exploration of Climates for Technical Updating and Performance," *Personnel Psychology,* Vol.40, pp.539-563(1987).

30. Kozlowski, S.W., & K.Hattrup, "A Disagreement About Within-Group Agreement : Disentangling Issues of Consistency Versus Consensus," *Journal of Applied Psychology,* Vol.77, No.2, pp.161-167(1992).

31. Landy, M.A., et al., "Intraclass Correlations : There's More There Than Meets the Eye," *Psychological Bulletin,* Vol, 93, No.3, pp.586-595(1983).

32. Landy, F.J., & J.L.Farr, "Performance Rating," *Psychological Bulletin,* Vol.87, No.1, pp.72-107(1980).

33. Lawlis, G.F.,& E.Lu, "Judgement of Counseling Process : Reliability, Agreement, and Error," *Psychological Bulletin,* Vol.78, pp.17-20(1977).

34. Levine, E.L., R.A.Ash, & N.Bennett, "Exploratory Comparative Study of Four Job Analysis Methods," *Journal of Applied Psychology,* Vol.65, pp.524-535(1980).

35. McCormick, E.J., P.R.Jeannett, & R.C.Meham, "A Study of Job Characteristics and Job Dimensions as Based on the Position Analysis Questionnaire(PAQ)," *Journal of Applied Psychology,* Vol.56, pp.347-368(1972).

36. Mitchell, S.K., "Interobserver Agreement, Reliability, and Generalizability of Data Collected in Observational Studies," *Psychological Bulletin,* Vol.86, No.2, pp.376-390 (1979).

37. Peterson, R.B., "The Interaction of Technological Process and Perceived Organizational Climate in Norwegian Firms," *Academy of Management Journal,* Vol. 18, pp.288-299(1975).

38. Pritchard, R.D., & B.W.Karasick, "The Effects of Organizational Climate on Managerial Job Performance and Job Satisfaction," *Organizational Behavior and Human Performance,* Vol.9, pp.126-146(1973).

39. Rothstein, H.R., "Interrater Reliability of Job Performance Ratings : Growth to Asymptote Level With Increasing Opportunity to Observe," *Journal of Applied Psychology,* Vol.75, No.3, pp.322-327(1990).

40. Sackett, P.R., & M.A.Wilson, "A Factors Affecting the Consensus Judgment Process in Managerial Assessment Centers," *Journal of Applied Psychology,* Vol.67, pp. 10-17(1982).

41. Schmidt, F.L., & J.E.Hunter, "Interrater Reliability Coefficients Cannot Be Computed When Only One Stimulus Is Rated," *Journal of Applied Psychology,* Vol. 74, No.2, pp.368-370(1989).

42. Selltiz, C., et al., *Research Methods in Social Relations,* Toronto : Holt, Rinehart and Winston(1959).

43. Schneider, B.&D.E.Bowen, "Employee and Customer Perceptions of Service in Banks : Replication and Extension," *Journal of Applied Psychology,* Vol.70, pp. 423-433(1985).

44. Shrout, P.E., & J.L.Fleiss, "Intraclass Correlations : Uses in Assessing Rater Reliability," *Psychological Bulletin,* Vol.86, No.2, pp.420-428(1979).

45. Tinsley, H.E., & D.J.Weiss, "Research Methodology," *Journal of Counseling Psychology,* Vol.22, No.4, pp.358-376(1975).

46. Winer, B.J., *Statistical Principles in Experimental Desigh(2nd ed.),* New York :

McGraw-Hall(1971).

47. Zedeck, S., & H.T.Baker, "Nursing Performance as Measured by Behaviroal Expectation Scal A Multitrait-Multirater Analysis," *Organizational Behavior and Human Performance,* Vol.7, pp.457-466(1972).

48. Zohar, D., "Safety Climate in Industrial Organizations: Theoretical and Applied Implications," *Journal of Applied Psychology,* Vol.65, pp.96-102(1980).